Computational Models of Perceptual Organization and Bottom-up Attention in Visual and Audio-Visual Environments

by

Sudarshan Ramenahalli

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

April, 2018

Abstract

Figure Ground Organization (FGO) - inferring spatial depth ordering of objects in a visual scene - involves determining which side of an occlusion boundary (OB) is figure (closer to the observer) and which is ground (further away from the observer). *Attention*, the process that governs how only some part of sensory information is selected for further analysis based on behavioral relevance, can be *exogenous*, driven by stimulus properties such as an abrupt sound or a bright flash, the processing of which is purely bottom-up; or *endogenous* (goal-driven or voluntary), where top-down factors such as familiarity, aesthetic quality, *etc,* determine attentional selection. The two main objectives of this thesis are developing computational models of: (i) FGO in visual environments; (ii) bottom-up attention in audio-visual environments.

In the visual domain, we first identify *spectral anisotropy* (SA), characterized by anisotropic distribution of oriented high frequency spectral power on the figure side and lack of it on the ground side, as a novel FGO cue, that can determine Figure/Ground (FG) relations at an OB with an accuracy exceeding 60%. Next, we show a non-linear Support Vector Machine based classifier trained on the SA features

achieves an accuracy $\approx 70\%$ in determining FG relations, the highest for a stand-alone local cue. We then show SA can be computed in a biologically plausible manner by pooling the Complex cell responses of different scales in a specific orientation, which also achieves an accuracy $\geq 60\%$ in determining FG relations. Next, we present a biologically motivated, feed forward model of FGO incorporating convexity, surroundedness, parallelism as global cues and SA, T-junctions as local cues, where SA is computed in a biologically plausible manner. Each local cue, when added alone, gives statistically significant improvement in the model's performance. The model with both local cues achieves higher accuracy than those of models with individual cues in determining FG relations, indicating SA and T-Junctions are not mutually contradictory. Compared to the model with no local cues, the model with both local cues achieves $\geq 8.78\%$ improvement in determining FG relations at every border location of images in the BSDS dataset.

In the audio-visual domain, first we build a simple computational model to explain how visual search can be aided by providing concurrent, co-spatial auditory cues. Our model shows that adding a co-spatial, concurrent auditory cue can enhance the saliency of a weakly visible target among prominent visual distractors, the behavioral effect of which could be faster reaction time and/or better search accuracy. Lastly, a bottom-up, feed-forward, proto-object based audiovisual saliency map (AVSM) for the analysis of dynamic natural scenes is presented. We demonstrate that the performance of proto-object based AVSM in detecting and localizing salient objects/events is in

agreement with human judgment. In addition, we show the AVSM computed as a linear combination of visual and auditory feature conspicuity maps captures a higher number of valid salient events compared to unisensory saliency maps.

Primary Reader: Prof. Ralph Etienne-Cummings

Secondary Reader: Prof. Mounya Elhilali

Acknowledgments

My sincere thanks to Prof. Ralph Etienne-Cummings, without whose guidance, it would be impossible to conduct this research. His advice and encouragement at the right time always helped clear my doubts, move forward toward completing the goals of my research, especially at a difficult time in my life. I wish to continue this collaboration in future and look up to him as a great source of inspiration. With deepest gratitude, I would like to thank Prof. Ernst Niebur for his guidance, kindness, patience, constant encouragement and support. He has been a source of inspiration and a great example, whom I aspire to follow in my own career. Without his advice, it would have been impossible for me to navigate the complex challenges of graduate school.

My sincere acknowledgement goes toward my committee members, Prof. Hynek Hermansky and Prof. Mounya Elhilali. My discussions with Prof. Elhilali have helped me immensely in getting an understanding of the challenges in auditory scene analysis. I would also like to thank Prof. Andreas Andreou for his support throughout my degree. His natural optimism and kind nature are the qualities I admire in him and

wish to inculcate myself. I would also like to thank Prof. Rudiger von der Heydt for sharing his knowledge and enthusiasm about border ownership and for many helpful discussions when I started out in this area. I cannot forget Dr. Stefan Mihalas in acknowledging his input to my work on local cues for figure-ground organization.

My thanks go to all the current and former lab members for creating a nice environment, lively discussions and critiques during the lab meetings. I would also like to acknowledge the help of ECE department staff and MBI staff. I cannot forget to thank my 7^{th} grade teacher, Sister. Lucy D'Souza for encouraging my interest in science and mathematics. Last, but not the least, my heartfelt thanks to my parents, sister and nephew without whose encouragement, it would have been very hard to live through the daily grind of grad school.

Dedication

This thesis is dedicated to my parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

We acquire information about the physical world around us through our sensory systems, process that information to form mental representation of the external world, learn based on that mental representation, which gives rise to intelligence. Understanding the mechanism of sensory information processing in the brain can help build intelligent machines. Among the many sensory systems, in this thesis, we focus on vision, audition; and their interaction.

The mechanism of sensory information processing can be broken down into three successive stages namely, Sensation, Perception and Cognition. Sensation, the process in which physical features of sensory stimuli are transduced into electrochemical signals, is the most primitive stage. In the perceptual stage, the electrochemical signals are processed in the brain by neural cells, called neurons to represent sensory objects and their relationship to each other in time and space. This mental representation is propagated from early visual/auditory cortical areas to higher areas in the form of neural spikes, or action potentials. Based on these mental representations, we develop higher level concepts and knowledge,

store them in memory, do logical inference *etc*, all relate to the final stage of information processing, called Cognition.

The work presented in this thesis can be roughly divided into two parts. The first part of the thesis (Chapters 2 - 5) is focused on neurally inspired algorithms related to perception, the second stage of sensory information processing, in the visual domain. Specifically, I develop computational models related to Perceptual Organization. The process [1, 2] of piecing together "bits and pieces" of visual information into environmental objects that we actually experience in the real world is termed as, *Perceptual Organization* (PO). The mutually complementary processes of Grouping and Figure Ground Organization (FGO) facilitate perceptual organization. Grouping [3, 4] refers to the mechanism by which the feature fragments are put together to form perceptual objects, which appear as distinct, coherent units of experience, segregated from the background. When such objects are viewed by an observer, depending on the observer's viewpoint or line of sight, some objects may become fully or partially occluded. In the case of partially occluding objects, Figure-Ground Organization refers to determining which side of an occlusion boundary (OB) is the occluder, closer to the observer, referred to as *figure* and which side is the occluded, far away from the observer, termed as *ground*. While PO deals with organizing raw sensory information into objects and events that we experience, *attention* is the process that governs how only some of that sensory information is selected for further analysis based on behavioral relevance. Attention and PO are tightly coupled [5], where one can influence the other depending on stimulus properties or behavioral needs. The second part of the thesis (Chapters 6, 7) is about neurally inspired computational models of audio-visual (AV) integration; specifically,

bottom-up attention (details to follow) in AV environments.

FGO is an inherently ill-posed problem as it involves representation of objects and their relationships in three-dimensional environment based on the two-dimensional projection of the image on retina. Biology and, to a lesser extent, Computer Vision has developed methods to deal with this difficulty heuristically. The task can be solved using a multitude of cues. If two stereoscopically related images of the scene are available and the image has appropriate internal structure (a suitable texture on both the occluding and the occluded object), disparity (the distance between corresponding image elements in the two projections) can be used to determine the absolute distances of both objects from the observer. If either of these conditions is not met, or to complement a result obtained from disparity measurements, FGO needs to rely on a variety of monocular cues.

Beginning nearly a century ago, Gestalt psychologists [1, 7, 8] established a number of "principles" or cues that determine Figure-Ground (FG) relationships[1]. These can be roughly subdivided into global and local cues, distinguished by the distance from the OB at which information is being collected to make the FG decision. Global cues such as symmetry [9], surroundedness [10], and size [11] of regions integrate information over a large spatial extent to determine FG relationship between objects. Local cues, on the other hand, achieve the same by analysis of only a small neighborhood near the boundary of an object. Some examples of local cues are T-Junctions [12] and shading [13], including extremal edges [14, 15].

[1]It is impossible to provide a historical background or an illustrated review on the vast topics of Grouping, FGO or Gestalt laws of PO within the scope of this thesis, hence it is not attempted. We recommend the excellent articles by Wagemans et al. [3, 4] for detailed reviews on these topics. We have introduced only the key concepts and terminologies necessary to understand and appreciate the work detailed in this thesis

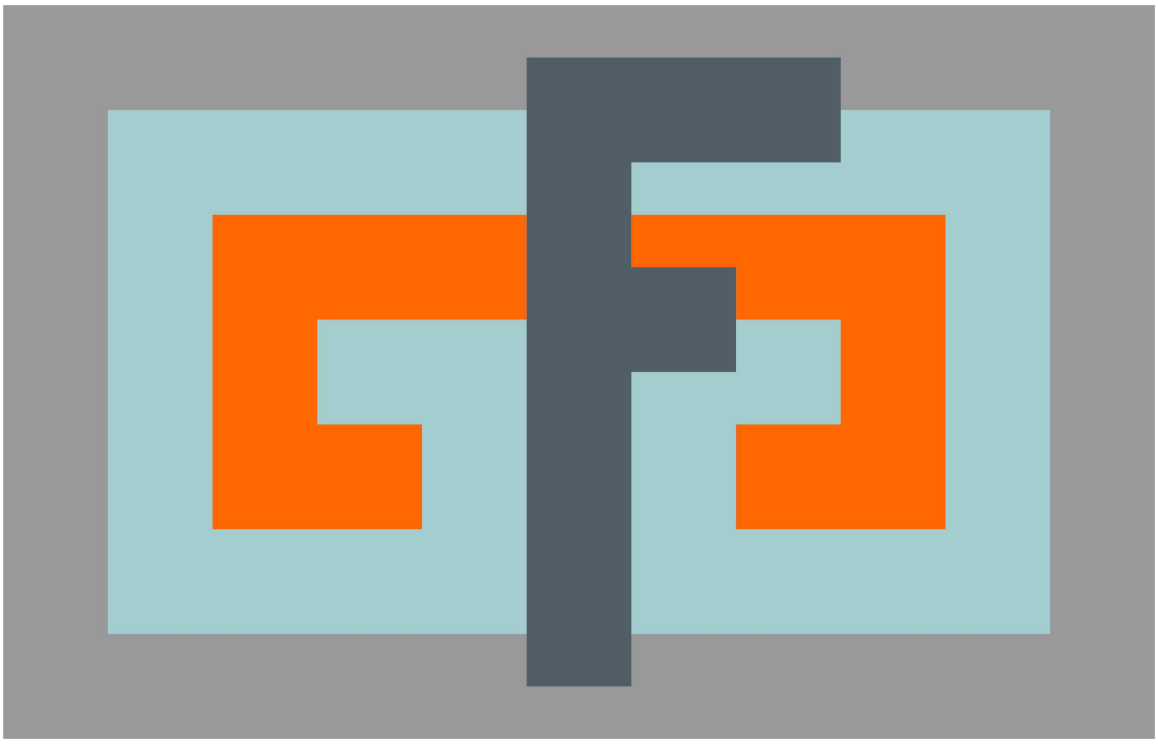

Figure 1.1: Figure ground organization with overlapping geometrical shapes. The letter F appears as figure and its boundaries are assigned to it ("owned by it"). In contrast, the ground regions do not own their borders and even if easily recognizable patterns are part of the background (see the light-colored letter G to the left of the F), they are found only through effortful scrutiny. Reproduced, with permission, from Qiu et al. [6]

Remarkably, in many cases FGO can be determined from very schematic, simplified visual "scenes" with nearly exclusively global cues. In more complex, realistic scenes, determination of figure ground relationships is not limited to global cues; instead, important contributions are made by local cues, too. Understanding how any or all of these local cues contribute to FGO is important for computational performance since, different from the global cues, they can be computed from a small number of pixels of the original image which can reduce computational complexity substantially.

Even if a single local cue only gives incomplete information by providing a bias for one or the other (binary) interpretation, combining many of these cues may "solve" the FGO problem in a statistical sense, and may require less computational resources than the use of global cues. Complex situations most likely require the integration of both global and local cues for a hybrid solution of the problem.

Understanding the importance of local and global cues in FGO, our goal in the first part of the thesis is to develop a neurally inspired computational model of FGO incorporating both local and global cues. With this end goal in mind, we first delve into finding new local cues of FGO, if any, by analyzing small image patches taken from both the sides abutting the OB. From this analysis, we identify a novel monocular FG cue that captures texture and intensity gradients present on the figure side of the OB, but absent on the background side. In the spectral domain, these gradients are characterized by anisotropic distribution of power in high frequency bins. Our method, which uses fast local 1D Discrete Fourier Transforms (DFTs) to characterize this new local cue, which we call *spectral anisotropy* (SA), is particularly suitable for machine vision applications, especially figure-ground la-

beling of an image. We then develop a biofidelitic algorithm for SA computation with the mathematical abstraction of the neural circuits typically found the visual cortex, the brain area where visual information is processed (See Section 1.1.1 for a brief overview of the anatomy and function of the visual system). We then incorporate this as one of the local cues in the neurally inspired model of FGO. Another type of local cues added in the FGO model are T-Junctions. The global cues that we incorporate in the FGO model are convexity, surroundedness and parallelism. We show, even the model without any local cues has an impressive performance and then compare the performances of the models with and without local cues. With the addition of local cues, we show, the model's ability to accurately determine FG relations improves substantially. We evaluate all the algorithms we develop on standard Computer Vision datasets, widely used in this type of work.

In the second part of the thesis (Chapters 6, 7), we study how spatial attention is controlled by the properties of the stimuli in AV environments. In daily life, attention is required for searching in a crowded scene (*ex,* face detection), tracking objects through occlusions, distinguishing speaker identities *etc,.* Attention can be *exogenous*, which is driven by the properties of the stimuli such as sudden onset of a loud sound or a bright flash, the processing of which is purely bottom-up; or *endogenous* (also called, goal-driven or voluntary), where top-down influences such as familiarity, aesthetic quality, *etc,* determine attentional selection. From a computational perspective, the most useful type of attention to study is the exogenous or bottom-up attention as stimulus properties can be more easily measured than top-down influences, which depends on the internal state of an observer.

Eye movements serve as a proxy for attention, in the sense, we can infer where attention

is being directed by tracking eye movements. Purely visual analysis of a scene begins with making fast eye movement, called "saccades" between fixation points. The fixation points generally fall on objects of interest to the observer. In AV environments eye movements are also influenced by the direction of sound in addition to the visual scene, which makes the analysis of AV scenes even more complex, which is why scientists and engineers have traditionally separated the analysis of an audio-visual scene into its constituent sensory domains. It was previously necessary to compartmentalize the analysis because of the sheer enormity of information as well as the limitations in experimental techniques and computational resources. With recent advances, it is now possible to perform integrated analysis of sensory systems including interactions within and across sensory modalities. A better understanding of interaction, information integration, and complementarity of information across senses can help us build many intelligent algorithms for scene analysis, object detection and recognition, human activity and gait detection, elder/child care and monitoring, surveillance, robotic navigation, bio-metrics *etc,* with better performance, stability and robustness to noise. Hence, we develop neurally inspired, bottom-up models of attention for automated scene analysis in AV environments.

There are many proposed models of attentional selection in the visual domain; *saliency* is one, which is stimulus driven or purely bottom-up[2]. The bottom-up saliency models [20, 21, 22] predict human eye fixations (attentional shifts) in images of natural scenes purely based on stimulus properties. One problem with these models is that the predicted fixation locations get stuck at sharp feature discontinuities, which typically lie between an object and

[2]Attention and related computational modeling have been active research topics for more than 40 years, producing an exhaustive amount of literature. We only introduce some important concepts required to follow the arguments in this thesis. For excellent reviews related to these topics, see [16, 17, 18, 19].

the background, whereas human eye fixations generally are centered on the objects. This is because saliency maps are based on simple center-surround differences of pixel intensities, so do not incorporate the notion of "objectness" of stimuli. To circumvent this problem, a notion of proto-objects or candidate objects or primitive objects is introduced in [23], further developed in [24, 25, 26]. Proto-objects are low-level, volatile object representations formed by bottom-up features before top-down attention is applied to form a clear representation of an object. The proto-object based saliency maps predict human attention better than simple pixel based saliency maps, but not close to human level. One such model is by Russell et al. [25], which groups feature fragments based on Gestalt principles of convexity, proximity, surroundedness and parallelism to construct proto-objects and compute saliency in the visual domain.

Efforts related to modeling attention in general, bottom-up saliency particularly, have so far been limited to individual [20, 21, 27, 28] sensory modalities rather than the combination of them. Slowly, interest in multisensory attention models, specifically audio-visual bottom-up attention is increasing. Here we propose to extend the concept of saliency map to both auditory and visual sensory modalities. First, we investigate the nature of multisensory interaction between the auditory and visual domains in a simple setting. More specifically, we consider the effect of a spatially co-occurring auditory stimulus on the salience of an inconspicuous visual target at the same spatial location among prominent visual distractors. Temporal concurrency is assumed between visual and auditory events. Based on the insights derived from this work, next, we extend our effort toward designing a proto-object based, feed-forward model of audiovisual saliency. The model consists of Color, Intensity,

Orientation and Motion as independent features in the visual domain, auditory location and loudness as features in the auditory domain. All the features undergo a grouping process to form proto-objects of each feature type. An audiovisual saliency map (AVSM) is computed from the proto-objects. We test the AVSM on real world data collected from an AV camera that can collect 360^0 audio and video that are temporally synchronized and spatially co-registered. The AVSM captures nearly all visual, auditory and audio-visually salient events, just as any human observer would notice in that environment.

1.1 Brief review of neural systems

In the next couple of sections, we briefly describe the anatomy and function of visual and auditory systems, which are referenced later in the thesis. We also touch upon the cross-modal anatomical pathways between visual and auditory systems to give an idea of the strong anatomical coupling between these two sensory modalities and the integrated nature of AV information processing in biological systems.

1.1.1 Anatomy and function of the visual system

The human eyes are approximately 6 cm apart, each having a vertical field of view (FOV) of 140^o, 60^o up and 80^o down. The horizontal FOV for both eyes (stereoscopic) is 114^o, but each eye can see an additional 40^o on its respective side, making the effective FOV approximately, 190^o. Light reflected from objects in the environment enters the through the small opening (Figure 1.2) called, *pupil* surrounded by *iris*, a type of muscular tissue, gets focused by the lens, the movement of which is controlled by ciliary muscles to accommodate

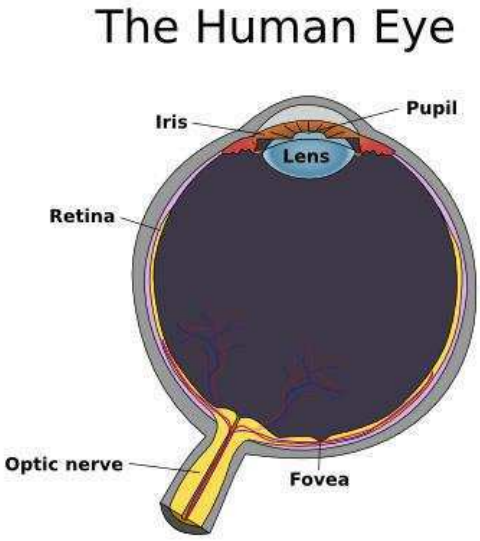

Figure 1.2: Anatomy of the human eye. Source [29]

for different object distances, to form a sharply focused image on the retina. The retina, the innermost surface of the eye made up of photoreceptors namely rods (sensitive to brightness) and cones (sensitive to color), acts like a screen on which the 2D image of the 3D world is projected. In simple terms, the mechanism of sensation at retina called, *photoisomerization*, can be explained as the transduction of light energy into neural signals by rods and cones. The sensitivity of the retina to light and cellular composition is not uniform all over. The central region of the retina where image of our gaze is formed (seen as a tiny depression in Figure 1.2), called *fovea* is very rich in cones whose density decreases away from the fovea non-linearly, while the rods are nearly absent in the fovea, but are more in number in the periphery. The visual acuity is highest in the fovea, which only covers about 2^o of the FOV (roughly about 1% of the retinal surface), hence foveal vision is necessary to perceive finer details.

At arm's length, the FOV of fovea is extremely small, roughly equal to twice the width

of a thumbnail [30], making it necessary for us to make saccades to bring different parts of the scene into focus as peripheral vision cannot provide detailed view. The peripheral vision is extremely helpful in detecting motion and navigating in the dark as rods present in the periphery are highly sensitive to light intensity (even a single photon releases neurotransmitters in rod cells [31]).

The transduced neural signal from retina reaches the inner most Retinal Ganglion Cell (RGC) layer. The RGCs have a "center-surround" receptive field (RF) structure, where two types are possible: (1) ON-center and (2) OFF-center.

The RF of a a neuron is defined as the sensory space within which when a stimulus of that sensory modality when presented alters its firing rate. Any change in stimulus property outside the RF has no effect on the neuron's firing properties. If a neuron is excitatory in nature, placing a suitable stimulus in its RF will increase its firing rate, where as if it is inhibitory, its firing rate diminishes. A center-surround RF is the organization of the RF in which the central disk region and the surrounding annular region have opposite sensitivities to light. An ON-center RF is excited (firing rate increases) when light is presented at the central disk, but inhibited when the surrounding annulus is exposed to light. On the other hand, the OFF-center RF has opposite property, *i.e.* , the cell is excited when no light falls on the central disk part and excited when it falls on the surround part. The center-surround RFs of RGC neurons are responsible for detecting contrast differences, which is an important step that facilitates detecting edges and boundaries in the primary visual cortex.

The neural signal from RGC is routed via the optic nerve to Lateral Geniculate Nuclei (LGN) and to the Superior Colliculi, structures in the thalamus (Figure 1.3), one in

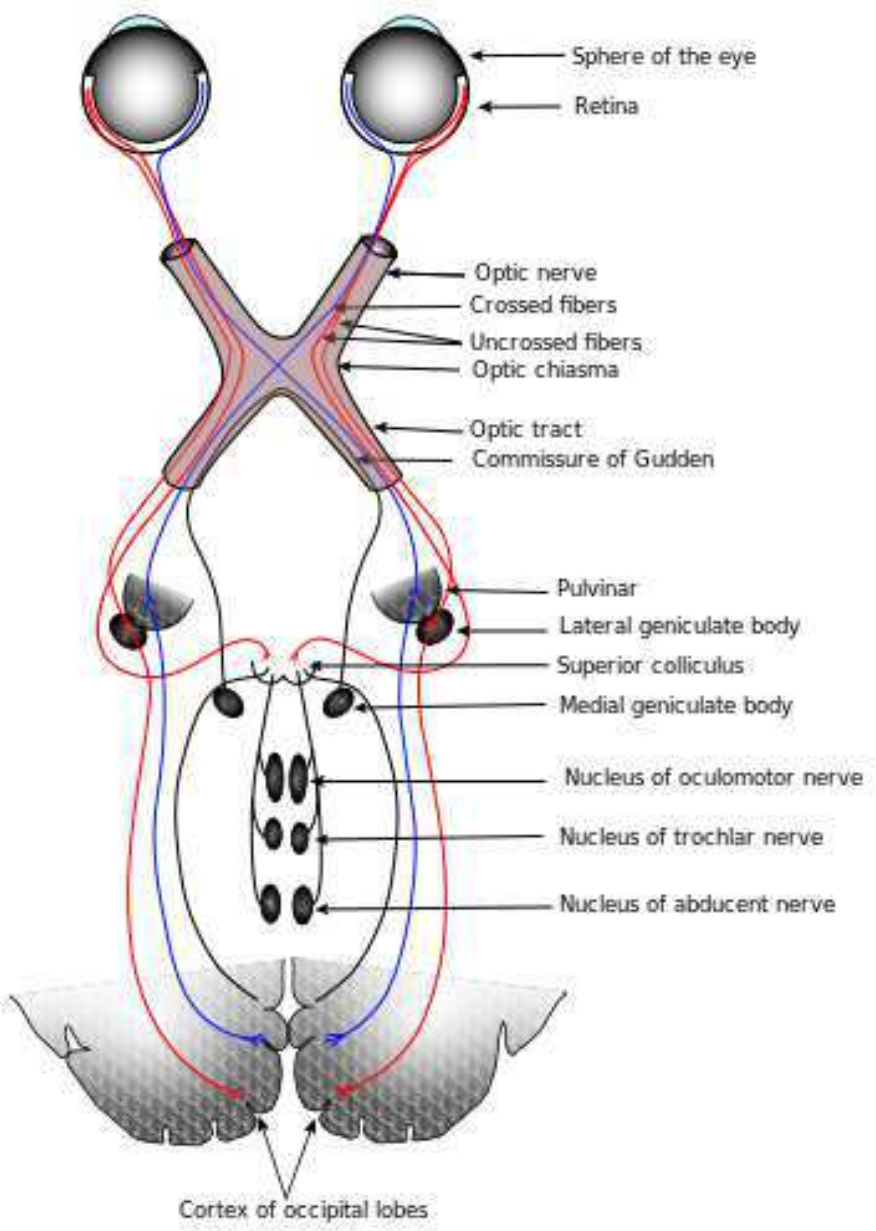

Figure 1.3: Pathway from retina to primary visual cortex. Source: [32]

each hemisphere of the brain. The LGN is functionally similar to RGCs having the same center-surround RF structure. Superior Colliculus (SC) is the center of convergence of many sensory (visual, auditory, somatosensory) inputs, hence recognized as the site of multisensory integration. It is also involved in controlling eye movements and sends afferent projections to higher areas of visual cortex, other than area V1 (discussed more in detail in Section 1.1.3). The neuronal fibers from the LGN project to primary visual cortex, also called striate cortex or area V1 (in primates) located in the occipital lobe. The fibers projecting from the left and right visual hemi-fields cross at optical chiasm before LGN and project to area V1 on the contralateral sides, respectively. So, stimuli on the right visual hemi-field stimulate the left hemisphere part of area V1 and *vice-versa*.

The neurons in area V1, as in other parts of the neocortex, are arranged in 6 horizontal layers, starting with layer 1 on the outer surface of the cortex and layer 6, being the innermost layer. Each layer receives feedback or sends feed-forward signals to different areas of the brain.

The neurons in area V1 have elongated RFs instead of the nearly circular RFs found in LGN. These cells respond to elongated bars whose orientation matches with the V1 neuron's preferred orientation. The neurons in area V1 can be classified as (i) Simple Cells, and (ii) Complex Cells depending on the response properties of cells. Both Simple and Complex cells respond to elongated drifting bars or gratings of their preferred orientation when they fall within their RFs. Simple cells have distinct, elongated ON and OFF subregions (also called, *lobes*), such that when a bar of preferred orientation and contrast falls on the respective sub-region, a neuronal spike is generated. As a result, the response of

Simple Cells is periodic for drifting gratings of preferred orientation. On the other hand, Complex cells do not have such ON and OFF sub-regions, as a result they respond to preferred orientation bars of either polarity giving rise to continuous spikes, rather than periodic spikes, as in the case of Simple Cells. Hence, Simple Cells are sensitive to contrast polarity of their preferred stimuli, whereas Complex cells are contrast polarity invariant. In addition, Complex cells are invariant to the exact position and phase of the stimulus, to some extent. Both Simple and Complex cells of area V1 are selective to orientation, spatial frequency, direction of motion, temporal frequency, disparity and color [33, 34]. As in the retina, primary visual cortex neurons also have selectivity to "red-blue", "green-yellow" and "white-black" opponent color channels [35]. This means, area V1 encodes the difference of the color pairs rather than individual colors.

Based on the spatial symmetry of the RF structure with respect to the imaginary central axis, Simple cells are further classified as Even and Odd symmetric cells. The Even symmetric cells have an elongated central excitatory sub-region flanked by inhibitory sub-regions on the sides or *vice-versa*. These RFs are symmetric about the center and resemble a damped cosine (symmetric, even function) waveform. The Odd symmetric cells have one half excitatory sub-region and the other half inhibitory sub-region and resemble a damped sine function, which is an odd symmetric function. Hence the RF of an Odd symmetric cell is 90^o phase shifted compared to that of an Even cell and *vice-versa*.

The neurons in area V1 have a *retinotopic* organization. In retinotopic organization, the neurons corresponding to adjacent locations in the visual field are anatomically adjacent to each other. Hence, topography is preserved all along the visual pathway, from retina to

LGN to area V1. In higher cortical areas post V1, the organization becomes more complex and the continuous first order retinotopic representation is lost. Retinotopic organization is also found in Superior Colliculus

Similar to the non-uniform distribution of cones and rods in the retina, even in the cortex more number of neurons are involved in processing foveal visual information compared to the periphery. This non-uniform representation of the visual field leads to a large proportion of the cortical surface dedicated to processing foveal information, in effect having smaller receptive fields, and very little cortical surface dedicated to peripheral visual field, hence processed by correspondingly larger RFs. This property of the cortical representation, measured by the amount of cortical surface area (in mm^2) dedicated to processing of information from one unit of visual field (measured in degrees), is termed, *cortical magnification*. Cortical magnification for fovea is higher than the periphery by a factor of about 100 [36].

As our work is not directly related to some other organizational principles of the striate cortex such as Orientation Columns, Ocular Dominance columns and Columnar Organization, they are not covered. Interested readers can refer Yantis [37] for more detailed information.

In addition to Simple and Complex cells, there are also hypercomplex or end-stopped cells [38] in the primary visual cortex, which have more recently been reclassified as subclasses of Simple and Complex cell types. The response of these cells depends, in addition to preferred orientation, also on the length of the stimulus, with the response decreasing with increase in the length of the stimulus. These can be Simple or Complex end-stopped

cells, also stopped at one or both ends. This distinct property is believed to play a role in curvature detection [38] and detection of corners or T-Junctions [39].

The area V1 sends feed-forward projections to area V2, an intermediate area in the visual processing hierarchy, specialized in detecting more complex patterns than area V1. Area V2 has been shown to have border ownership preference [40], a property because of which V2 neurons fire vigorously only when an object is placed on its preferred side at its preferred orientation, but do not respond for objects placed on their non-preferred side even though the local content within their RF is same in both cases. Like V1 neurons, neurons in V2 also have orientation and spatial frequency selectivity, in addition to selectivity for binocular disparity and the orientation of illusory contours. It has also been shown that neurons in V2 are selective for combination of V1 features such as curvatures [41, 42], corners, texture [43] and other complex second order features [44] found in natural scenes. Like other areas of the visual cortex, there are feedback connections from V2 to V1. Reciprocal connections between V2 and V4 in the ventral part and between V2 and area MT (middle temporal) in the dorsal part of the cortex are found.

The neurons in area V4, which receive input from V2, have much larger RFs compared to V2, are selective to color and shape features of intermediate complexity and also curvature [45, 46]. Also, neurons in this area are substantially modulated by attention, hence it serves as an interfacing stage between top-down and bottom-up processing. Neuronal projections from V4 go to the Inferotemporal (IT) cortex, which is involved in recognition and identification of more complex shapes [47]. The cells in this area have very large RFs, exhibit translation, contrast, size, color and rotation invariance and are modulated by at-

tention [48, 49]. The ventral visual pathway consisting of V1, V2, V4 and IT, specializing in processing the shape and color of objects, helping us identify the objects, is generally referred to as the "what" pathway.

The neural signal pathway on the dorsal part consisting of V1, V2 and middle temporal (MT) area specializing in locating objects in space, their motion direction, speed, *etc* in egocentric coordinates, is called the "where" pathway. Both V1 and V2 send afferent projections to area MT. MT makes connections with many other areas. As this is not the main focus of our study, we will not go into the details. Interested readers may refer to [50].

Even though the "two-stream" hypothesis of ventral stream specializing in object form processing and dorsal stream specializing in motion and spatial location processing is predominant, the dissociation between the two streams is not as simple as was previously thought [51, 52].

1.1.2 Anatomy and function of the auditory system

In this section, only parts of the auditory system relevant in the context of the thesis are described. Since the thesis does not primarily deal with the processing of audio separately, but in the context of audiovisual integration, those anatomical parts and functions that are relevant in the context of crossmodal processing are focused more.

The human auditory system is sensitive to sound in the range 20 Hz - 20,000 Hz. Sound hitting the ear is funneled into the narrow auditory canal by the pinna (Figure 1.4). The shape of the pinna is different for each individual and has the effect of filtering the sound signal and amplifying certain frequencies, which depends on the shape of the pinna

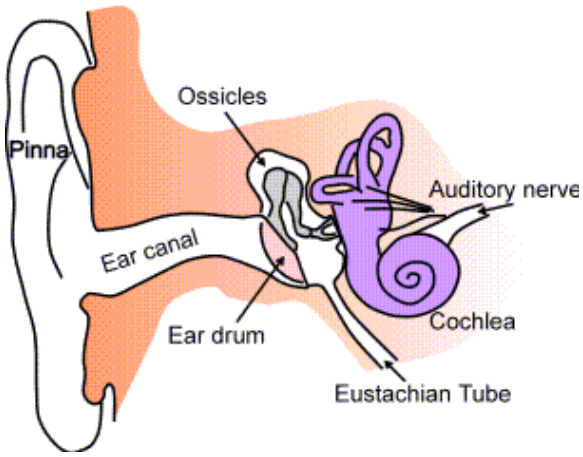

Figure 1.4: Anatomy of the human ear. Source: [54]

and direction of sound. Hence, when reproducing spatial sound for earphones it becomes important to take this into consideration [53].

The auditory canal amplifies sound in the range of 3 kHz - 12 kHz, which is then conveyed to the ear drum or tympanic membrane, which vibrates in response to the sound waves. The vibrations are amplified by the three bones called, *ossicles* and fed into the *cochlea*, a snail shaped hollow tapering coiled tube filled with a fluid, separated by the basilar membrane (Figure 1.5) into upper and lower chambers . The basilar membrane, thick and narrow at the base, closer to the tympanic membrane is most responsive to high frequencies near 20 kHz. The sensitivity of the basilar membrane varies inversely with the frequency along its length, which is sensitive to frequencies as low as 20 Hz at the other end of the membrane, which is thin and wider. The function of basilar membrane is similar to that of a filter bank with multiple overlapping bandpass frequencies. The hair cells situated on the basilar membrane transduct the mechanical vibration of the basilar membrane into neural signals and amplify them which are routed via the auditory nerve to the cochlear

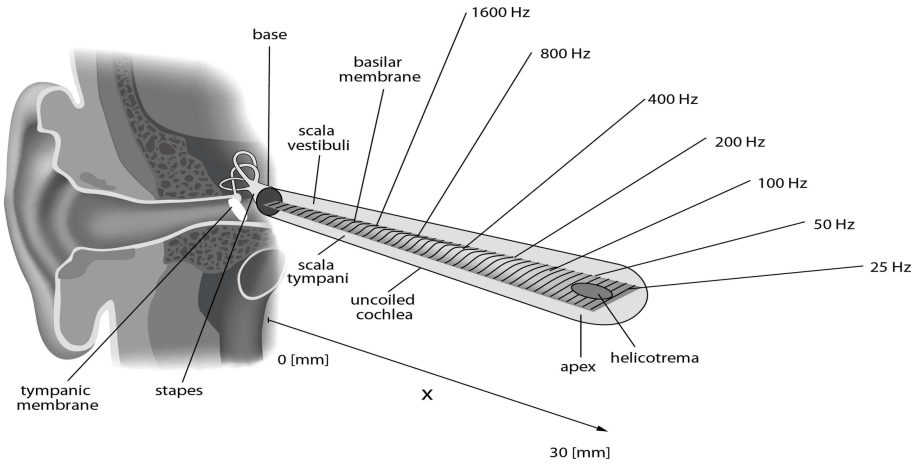

Figure 1.5: Characteristic frequencies of the basilar membrane, the frequencies to which different parts of the cochlea are most sensitive. Source: [55]

nucleus in the brain stem. Not only there is a "place code" or "tonotopic organization" for frequency representation in the cochlea, wherein different places along the basilar membrane are most selective to their preferred or "characteristic frequency", but there exists also a temporal code for the same.

Since neural firing rates are limited to a few hundred spikes per second, the auditory system uses temporal coding mechanism for representing high frequencies above, say a 1000 Hz, where the auditory nerves fire in synchronization with the peaks of the auditory signal, even though not for every single peak in the auditory signal, but always in phase with the peaks of the incoming signal. In addition to the frequency, auditory system also codes for the amplitude of the auditory signal. Similar to the way visual system has to deal with light intensity, the auditory system has to deal with large dynamic range of sound amplitude or loudness, measured in decibel sound pressure level (dB SPL), which ranges from 20 dB SPL (quiet room) to 120 dB SPL (threshold of hearing). This is accomplished by the auditory system by employing many neurons in proportion with the loudness.

Both feed-forward (ascending) and feedback (descending) auditory pathways exist between cochlea and the cortex. While the function of descending pathway is not yet fully understood, its main role is found to be modulatory, activation of acoustic reflex to protect ear from damage and in inhibiting the irrelevant sounds in order to better attend to the relevant auditory stimuli [56]. The ascending pathway can be divided into primary and secondary pathways (Figure 1.6). Neural signals from Cochlea are conducted to the Cochlear Nucleus in the brain stem through the auditory nerve. From the Cochlear Nucleus, auditory nerves that project to structures in the brain stem, mid brain, thalamus and cortex on the opposite side (contralateral) form the primary auditory pathway, where as the projections from Cochlear Nucleus to structures on the same side (ipsilateral) form the secondary pathway. In the primary pathway, we see direct projections from the Cochlear Nucleus to the contralateral Trapezoid Body (involved in sound localization [57]) and Superior Olivary Complex (neural signals from both ears converge here, responds to binaural signals) in the brain stem and Inferior Colliculus (IC) in the mid brain. The Trapezoid Body projects again to the Superior Olivary Complex (involved in detection of inter-aural level and time differences, mechanisms of sound localization [58]), which in turn projects to IC. From IC, neural signals travel to the auditory cortex via Medial Geniculate Body (MGB) in the Thalamus. The MGB receives secondary input from the ipsilateral IC. The secondary auditory pathway consists of projections from the Cochlear Nucleus to the ipsilateral Superior Olivary Complex, IC, MGB and auditory cortex. Another projection to the MGB comes from contralateral IC, which is also routed to the auditory cortex.

The Inferior Colliculus together with the Superior Colliculus located rostrally form

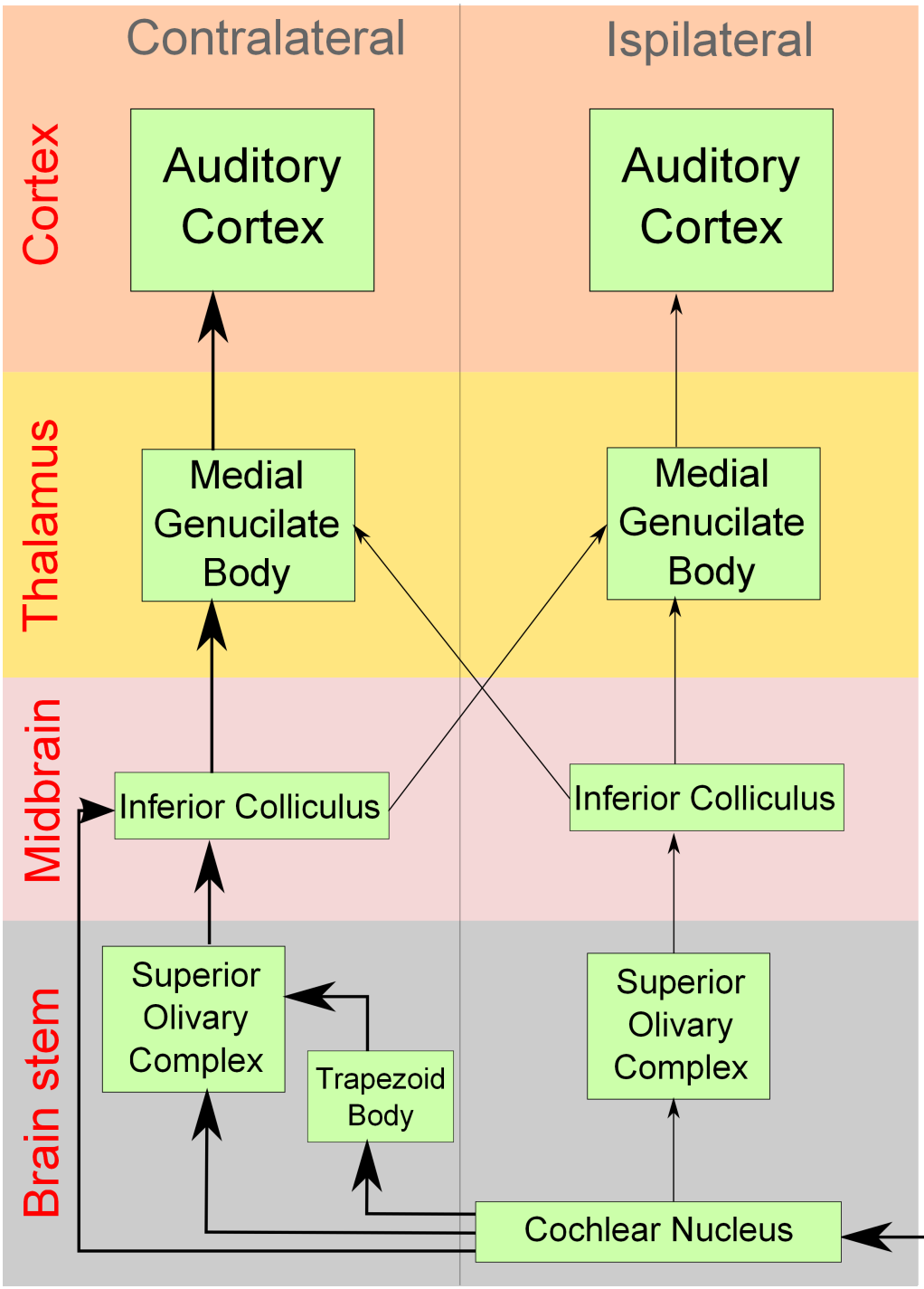

Figure 1.6: The primary (thick lines) and secondary (thin lines) auditory pathways from Cochlea to the auditory cortex. The Cochlear Nuclei receive input via the auditory nerve from the cochlea on the same side (cochlea not shown in figure)

the multisensory integration sites within the mid-brain (more on this in Section 1.1.3). Auditory-somatosensory integration takes place in IC, in addition to spatial sound localization through binaural hearing. The MGB is the thalamic relay between IC and auditory cortex. Tonotopic organization is maintained all along the auditory pathway. The auditory cortex located inside the Sylvian fissure of the temporal lobe consists of the primary auditory cortex or A1, the rostral core and rostrotemporal core. Together, the three areas constitute the auditory core region which is wrapped around by the belt and parabelt regions (Figure 1.7). In all the three areas there is tonotopic organization of neurons with neurons sensitive to low frequencies at one end, and high frequencies at the other end. The neurons in cortical as well as sub-cortical areas have different tuning characteristics to frequency, some are narrowly tuned and others are broadly tuned. While the area A1 responds to pure tones much more vigorously, the belt and parabelt regions are tuned to more complex features of sound required for recognizing and discriminating different auditory objects. Similar to visual cortex, the auditory cortex also has distinct regions for processing the identity ("what") and location ("where") of sound sources. The "what" pathway extends from the core region anteriorly toward the belt and parabelt areas and and further beyond in the temporal lobe, whereas the "what" pathway extends posteriorly concluding in the posterior parietal cortex (PPC) [59]. Again, similar to the visual system, this distinction of "where/what" pathways is not very strict and there is massive amount of convergence of inputs from different sensory modalities at all levels of the cortical processing hierarchy, which is the focus of the next section.

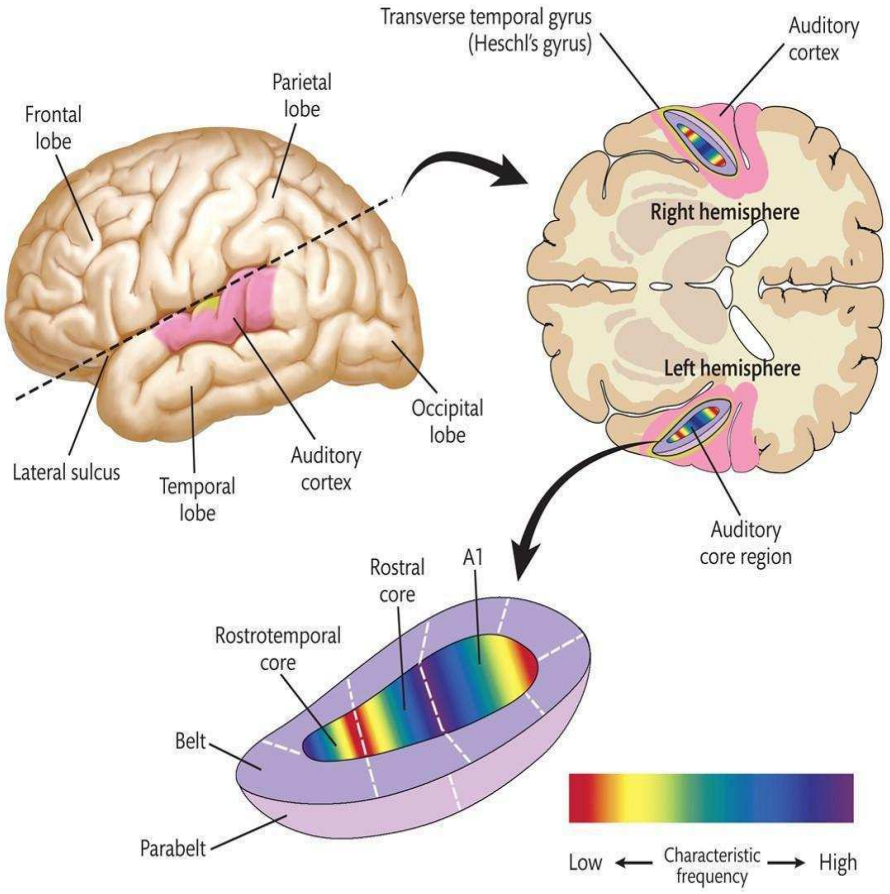

Figure 1.7: The auditory cortex with auditory core, belt and parabelt regions. Tonotopic organization is maintained in all three areas. Reproduced with permission from [37]

1.1.3 Cross-modal anatomical pathways

The traditional view that sensory signals are processed in dedicated areas of the brain in a hierarchical fashion and merged later at higher levels of cortex is becoming increasingly obsolete [60, 61, 62]. Recent evidence from physiology, psychophysics and imaging studies suggests cross-modal connections exist even in the lowest cortical areas as well as in the thalamic areas. Even though connections exist between visual, auditory and somatosensory modalities, here we restrict ourselves to the crossmodal connections between the visual and auditory areas only.

Crossmodal connections exist between visual and auditory areas in the cortex, thalamus and mid-brain. The connections can be feedback or feedforward. The primary site of multisensory integration is identified as Superior Colliculus (SC) in the cat, located in the mid-brain [61, 63, 64, 65]. The SC is involved in orienting behavior and localization of events. The superficial layers (layers 1 – 3) are unisensory, responsive to visual stimuli only, while the deep layers (layers 4 – 6) are multisensory in nature, responsive to all combinations of auditory, visual and somatosensory stimuli. The SC has topographic maps of different sensory modalities which are spatially registered. The response of SC neurons varies from super-additive (for weak multisensory stimuli) to additive to sub-additive (for strong stimuli) as the stimulus intensity or signal-to-noise ratio increases. This inverse relationship between response strength of SC neurons and the stimulus strength is termed, "inverse effectiveness". In addition to feedforward visual inputs from LGN and auditory inputs from the IC, SC also receives feedback inputs from the association cortices, which are crucial for multisensory integration. The SC has bi-directional connections with Pulvinar,

which is also connected to pre-motor areas, V1 [66] and MT [67]. The frontal eye field (FEF), which is responsible for generating saccades, gets input from SC via thalamus and responds to both visual and auditory stimuli relatively fast (20 – 85 ms) [68]. The spatial RFs of SC neurons are similar to their unisensory counterparts [69]. While some SC neurons have a single dominant "hot-spot" region in their RF where they are maximally responsive to multimodal stimuli, others have several spatially discrete "hot-spots".

The cross-modal anatomical connections between auditory and visual cortices exist starting from earliest unisensory areas, all the way up to the association areas (Figure 1.8). There are direct projections from the auditory cortical association areas (belt and parabelt regions) to the area V1 [70], V2 [71] and Prostriata, a region between V1 and V2, containing peripheral field representation [72] and involved in dorsal stream processing. These projections mainly serve to represent the peripheral visual field, very few neurons from auditory cortex send projections to foveal visual cortex. Very few projections were found from the area A1 to visual areas [73], but this could be due to a limitation of the study technique. On the other hand, projections from the areas V2 and Prostriata extend to caudal parabelt area (CPB), belt and Temperoparietal (Tpt) areas. All these cross-modal projections are found to have a modulatory effect rather than driving the neurons in the other sensory area. No direct projections from area V1 to auditory cortex are found so far.

The advantage of such cross-modal connections are many. The response to a stimulus in the auditory core area (A1) is very fast with a delay of only 10 ms compared to the visual area V1, where the delay is $\approx$ 30 ms. But, the spatial accuracy in visual domain is much higher compared to auditory domain. Hence, audiovisual integration at early level brings

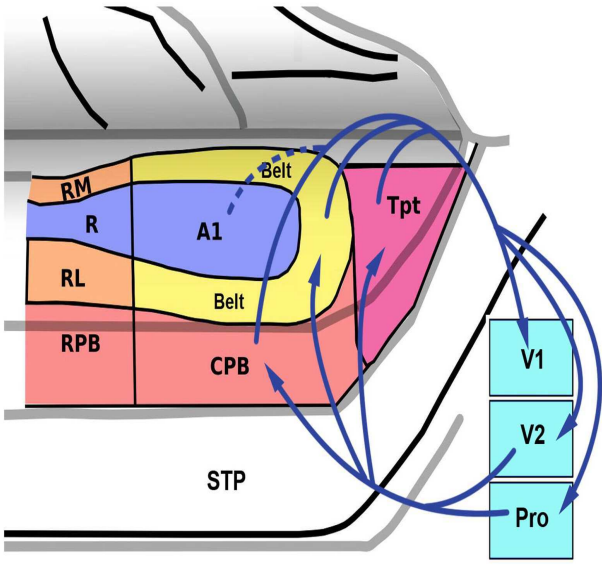

Figure 1.8: Cross-modal anatomical connections between early auditory and visual areas. There are bi-directional projections between primary and secondary visual cortices to primary auditory cortex and auditory association areas (belt and parabelt regions). The projection from A1 to visual cortex is not significant, but exists (dashed lines). These connections serve mainly to represent peripheral field. There is very little direct interaction between the two sensory cortices for space corresponding to foveal visual field. Tpt: Temperoparietal area; CPB: Caudal parabelt area; Pro: Prostriata; RPB: Rostral parabelt area. Reproduced with permission from [72]

about the best of both worlds: the accuracy of visual domain paired with the speed of the auditory domain. As a result of integration, we observe the response latency of area V1 to reduce. Also, the short latency observed during audiovisual integration in the traditional unisensory areas shows that the multisensory effects observed are not primarily because of a feedback, but due to direct heteromodal connections.

Coming to higher areas in the hierarchy, most neurons in the frontal, parietal and temporal lobes are multisensory, which is now well established. The prefrontal cortex receives information from both visual and auditory cortices. The premotor cortex can also be regarded as a site of multisensory integration as both unisensory inputs from respective association areas and multisensory inputs from parietal lobe converge there.

The Anterior Ectosylvian Sulcus (AES) located at the border between frontal, parietal and temporal lobes has multisensory neurons with spatially registered RFs and exhibits the typical properties of multisensory integration. The RFs of AES neurons are substantially larger than SC neuron RFs. In general, the RFs of multisensory neurons in higher areas are heterogeneous, large and do not have a very strict spatiotopic organization. The areas Lateral Intraparietal area (LIP) and Ventral Intraparietal area (VIP), buried inside the Intraparietal sulcus (IPS) receive inputs from auditory, visual, vestibular and motor areas [74]. The temporal part of Superior Temporal Sulcus (STS) is connected to auditory belt and parabelt areas and occipital visual areas, which provides the neurons in the Superior Temporal Place (STP), multisensory properties. The STP in turn projects to V1 providing feedback.

Lastly, it is important to emphasize the thalamocortical audiovisual pathways that

play an important role in fast communication of information between cortex and thalamus and also between distant cortical areas. Corticocortical information transmission is slow. However, fast communication is achieved between distant cortical areas through special connections via several thalamic nuclei. For example, pulvinar nucleus which is connected to many cortical areas acts like a hub for fast interactions between cortical areas.

1.2 Motivation

One of the motivations of this thesis is to understand the working of neural circuits that give rise to perception with the goal of building useful artificial intelligence algorithms applicable in real world natural scenes. Even though it is possible to understand *where* and *when* in the brain sensory information processing takes place using techniques such as functional Magnetic Resonance Imaging (fMRI), Encephalogram (EEG) and single cell electrophysiology, what is important from the perspective of neurally inspired algorithm design is to understand *how* it happens.

With single cell electrophysiology, it is possible to study how a single neuron or a few neurons processes information, but a good deal of the meaningful activity takes place in networks, which consist of say, 100 - 1000 neurons. Such networks, generally termed neural populations are an abstraction, in which, we need to understand the mechanism of sensation and perception, in order to design algorithms. At present computational modeling is an elegant way to elucidate the mechanism of *how* sensory information is processed in these neural populations. Computational modeling can inform us about the type of neural population activity that gives rise to the observed behavior in imaging or physiology

experiments. Hence, we take this approach to understand sensory information processing mechanism in neural populations, with a firm foundation of anatomical connections.

With sensors such as cameras, microphones, gyroscopes, *etc* becoming increasingly common and inexpensive, our ability to store, compress and transmit data is also increasing at a rapid rate. So, it is now possible to acquire and analyze vast amounts of data in order to optimize or automate any type of application. On the other hand, such automated analysis of sensory data is becoming increasingly important for a variety of robotics and surveillance applications such as autonomous driving, remotely monitoring infants, elder care, quality assurance, human computer interaction, *etc*.

Motivated by the need for intelligent algorithms for automation, we proceed to develop neurally inspired computational models of perceptual organization and bottom-up attention. Why neurally inspired? Because the human brain is the most sophisticated and intelligent system we have known so far. Even a two year old child can do complex reasoning about objects and their relationships in its environment which is not yet achievable, even by the most advanced computers. Besides, approaches using machine learning [75] (ML) and deep learning [76, 77, 78] (DL) have some drawbacks. ML/DL systems are like a "black box", causal analysis of their behavior is not straight-forward. With neurally inspired computational algorithms with anatomical and functional underpinnings, each component has a distinct purpose. So, we can understand *why* the system behaves in a certain way, and as a result, add or remove components to achieve desired behavior. Hence, we believe, taking an approach that is closely modeled after biological vision/audition can lead to better systems.

In the visual domain, there has been significant amount of research on computational algorithms related to low level visual processes such as edge detection, image segmentation, *etc*. In the same way, higher level visual processes like object detection, recognition, *etc* have also received considerable attention in the Computer Vision community. But, computational algorithms for FGO, a mid-level visual phenomenon, has received little attention. With growing Computer Vision applications the importance of FGO is becoming evident in the industry. For example, to automate points counting in a basket ball game, we need to verify whether the ball made the basket, the player was standing behind the three-point line or not, all require occlusion reasoning or determining FG relations. Similarly, in autonomous driving and many other industries, the need for FGO is becoming strongly evident. Moreover, research in the field of electrophysiology on the neural mechanism of FGO has produced many useful and interesting breakthroughs that we can incorporate into computational algorithms.

The neural mechanism by which FGO is achieved in the visual cortex is an active area of research, referred to as Border Ownership (BO) coding. Remarkably, von der Heydt and his colleagues [6, 79, 80] discovered that the activity of individual cells in primate extrastriate cortex represents the BO relationships that likely underlie FG organization. Neuronal recordings in macaque monkeys show that the majority (about two-thirds) of orientation selective cells in area V2, the second largest visual area in macaque and the largest in humans, is BO selective [79, 80]. The visual stimuli in these experiments were devoid of most local cues, in fact all with the exception of a small number of L and T junctions, and even these were far away from the classical receptive fields of the recorded neurons.

von der Heydt and his collaborators found that neurons in early and intermediate visual areas nevertheless represented the global structure of the scene. This was the case even though the classical RFs of the recorded neurons only covered a very small fraction of the perceived figure and ground elements. Recently, Williford and von der Heydt [81], even more remarkably show for the first time, that V2 neurons maintain the same BO preference properties even for objects in complex natural scenes.

Motivated by the neuroscientific advances and the need for FGO algorithms in the industry, we decided to build neurally inspired, computational models of FGO. The low computational cost and proven usefulness of SA, based on our own studies led us to add SA as a local cue into the model. T-Junctions are again local cues, which are generally regarded as one of the strongest occlusion cues, hence we decided to add them. The computation of both cues can be explained on the basis of cells found in the visual cortex, which further strengthened our motivation.

As natural environment and our interaction with it is essentially multisensory, where we deploy visual, tactile and/or auditory senses to perceive, learn and interact with our environment, integrated analysis of multisensory information naturally leads to a rich understanding of a situation. Also, we can now record spatial audio and video of any desired FOV making the analysis of spatial attention in AV environments possible. Plus, there is a huge need for such algorithms in surveillance. Manually monitoring hundreds of cameras is labor intensive, time consuming and demands constant human vigilance. Human attention is limited and we may miss some important events when we are paying attention to something else. This is a well studied phenomenon, as demonstrated by the Invisible Gorilla

Experiment [82].

Recent evidence from neuroscience [60, 83] suggests that the traditional view that the low level areas of cortex are strictly unisensory, processing sensory information independently, which is later on merged in higher level associative areas is increasingly becoming obsolete. This has been proved by many fMRI [84, 85], EEG [86] and neuro-physiological experiments [87, 88] at various neural population scales. There is now enough evidence to suggest an interplay of connections between thalamus, primary sensory and higher-level association areas which are responsible for audiovisual integration. The broader implications of these biological findings may be that learning, memory and intelligence are tightly associated with the multi-sensory nature of the world.

Motivated by the multisensory nature of our environment, technological advances in spatial audio, advances in neuroscientific research into audiovisual integration and the need for such algorithms, we decided to build bottom-up audiovisual saliency algorithms. Moreover, computational modeling efforts related to multisensory saliency models have been historically limited. With the availability of audiovisual recording technology, which is getting better and cheaper with time, the properties of the stimuli in the environment can be measured easily and more accurately than the internal state of the observers. On the other hand, top-down influences like internal state, prior experiences, goals, *etc* can neither be probed nor quantified. Moreover, these top-down factors vary widely across observers. Since bottom-up saliency is purely based on stimulus properties, which are measurable, largely remain constant across observers, we were naturally motivated to develop bottom-up AV saliency models.

1.3 Contributions of the thesis

In the visual domain, with the goal of building a model of FGO incorporating local and global cues, we first focus on a novel set of local cues in the intensity patterns along OBs which we show to differ between figure and ground. Image patches are extracted from natural scenes from two standard image databases along the boundaries of objects and spectral analysis is performed separately on figure and ground. On the figure side, oriented spectral power orthogonal to the occlusion boundary significantly exceeds that parallel to the boundary. This anisotropic distribution of oriented high frequency spectral power on the figure side is called, Spectral Anisotropy (SA). This SA is present only for higher spatial frequencies on the figure, and absent on the ground side. The difference in SA between the two sides of an OB predicts which is the figure and which the background with an accuracy exceeding 60% per patch. We also show, SA of close-by locations along the boundary co-varies but is largely independent over larger distances which allows to combine results from different image regions. Given the low cost of this strictly local computation, SA along OBs is a valuable cue for FGO. A data base of images and extracted patches labeled for figure and ground is made freely available. Our findings related to this research were published in the journal, Vision Research [89], and at other venues [15, 90].

Next, we show a non-linear Support Vector Machine based classifier trained on the Spectral Anisotropy features derived from image patches extracted along the OB achieves an accuracy near 70% per local patch on the task of deciding which side of an OB is the foreground. This is the highest Figure-Ground classification accuracy for a stand-alone local cue reported so far, exceeding other cues such as, convexity, lower region, *etc.* Our

findings were published in the IEEE Conference on Information Sciences and Systems (CISS 2012) [91].

We then show computation of SA in a biologically plausible manner is possible by pooling the Complex cell responses at different scales in a specific orientation. The biologically plausible computation also achieves an accuracy exceeding 60% on the task of deciding which side of an OB is figure for all boundary locations in the ground truth maps of BSDS figure/ground database. These results firmly establish SA as a novel and valid local cue of FGO and its biological plausibility.

After establishing SA as a valid figure-ground cue, we present a biologically motivated, feed forward computational model of FGO incorporating convexity, surroundedness, parallelism as global cues and SA, T-junctions as local cues, where SA is computed in a biologically plausible manner. The model consists of three independent feature channels, Color, Intensity and Orientation, but SA and T-Junctions are introduced only in the Orientation channel as these properties are specific to that feature of objects. We evaluate model performance based on figure-ground classification accuracy (FGCA) at every border location using the BSDS 300 figure-ground dataset. FGCA for an image is defined as the percentage of the total number of boundary pixels in the ground truth figure/ground label map for which a correct figure/ground classification decision is made by the model. We show that each local cue, when added alone, gives statistically significant improvement in the FGCA of the model suggesting its usefulness as an independent FGO cue. The model with both local cues achieves higher FGCA than the models with individual cues, indicating SA and T-Junctions are not mutually contradictory. Compared to the model with no local cues, the

feed-forward model with both local cues achieves $\geq 8.78\%$ improvement in terms of FGCA. The manuscript is to be submitted to a suitable journal for publication.

In the audio-visual domain, first we build a simple computational model to explain how visual search can be aided by providing concurrent, co-spatial auditory cues. Our model shows that adding a co-spatial, concurrent auditory cue can enhance the saliency of a weakly visible target among prominent visual distractors, the behavioral effect of which could be manifested as faster reaction time and/or better search accuracy. The computational model explains the results of some previous psycho-physics experiments [92, 93]. This work was published in the 2013 IEEE CISS [94].

The final contribution of the thesis is a bottom-up, feed-forward, proto-object based audiovisual saliency map (AVSM) for the analysis of dynamic natural scenes. A specialized audiovisual camera with 360° Field of View, capable of locating sound direction, is used to collect spatiotemporally aligned audiovisual data. We demonstrate that the performance of proto-object based AVSM in detecting and localizing salient objects/events is in agreement with human judgment. In addition, the proto-object based AVSM that we compute as a linear combination of visual and auditory feature conspicuity maps captures a higher number of valid salient events compared to unisensory saliency maps. The results will be submitted to a suitable conference.

A list of publications resulting from the work presented in this thesis is given below:

[1] S. Ramenahalli, S. Mihalas, and E. Niebur. Figure-ground classification based on spectral properties of boundary image patches. In *IEEE CISS-2012 46th Annual Conference on Information Sciences and Systems*, pages 1–6, Princeton, NJ, 2012. IEEE Information Theory Society.

[2] Sudarshan Ramenahalli, Stefan Mihalas, and Ernst Niebur. Spectral inhomogeneity provides information for figure-ground organization in natural images. *Society for Neuroscience Annual Meeting*, 2011.

[3] Sudarshan Ramenahalli, Daniel R Mendat, Salvador Dura-Bernal, Eugenio Culurciello, E Nieburt, and Andreas Andreou. Audio-visual saliency map: overview, basic models and hardware implementation. In *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, pages 1–6. IEEE, 2013.

[4] Sudarshan Ramenahalli, Stefan Mihalas, and Ernst Niebur. Local spectral anisotropy is a valid cue for figure–ground organization in natural scenes. *Vision research*, 103: 116–126, 2014.

[5] Sudarshan Ramenahalli, Ernst Niebur, and Ralph Etienne-Cummings. A model of figure ground organization incorporating local and global cues, (to be submitted to a journal).

[6] Sudarshan Ramenahalli, Ernst Niebur, and Ralph Etienne-Cummings. A proto-object based audiovisual saliency map, (to be submitted to a conference).

1.4 Thesis organization

This thesis is organized as follows:

Chapters 2 - 5 cover the first part of our work in the visual domain related to Figure Ground Organization. The second part of the thesis, related to bottom-up attention in audio-visual environments, is detailed in Chapters 6 and 7.

In Chapter 2, we develop a simple and efficient algorithm to detect Spectral Anisotropy on figure and ground sides of an OB using averaged, 1D, oriented Fourier power spectra. We show SA is a useful and valid FGO cue. Then, with co-variance analysis, we show that SA of close-by locations along the boundary co-varies but is largely independent over larger distances. We also show SA based figure/ground assignment is robust to variation in patch sizes, image sizes, *etc* in Appendix A.

In Chapter 3, we train a non-linear Support Vector Machine classifier with radial basis function (RBF) kernels using Spectral Anisotropy features to achieve near 70% accuracy in determining the figure-ground relations.

For the work done in Chapters 2 and 3, we created a database of small images patches extracted along the OB from BSDS 300 and MIT LabelMe images, made available at `http://cnslab.mb.jhu.edu/data/codeanddata.html`

The biologically plausible computation of SA by pooling Complex cell responses of a certain orientation, but multiple scales, is covered in Chapter 4. We show the biologically

plausible SA computation gives results similar to what we observe in Chapter 2, even when computed on all boundary locations labeled for figure/ground ground truth in the BSDS figure/ground dataset, instead of randomly selected patches.

Chapter 5 is about the model of Figure Ground Organization incorporating both local and global cues. We show adding local cues results in statistically significant improvement in the FGO model's performance.

Chapter 6 describes a model of audiovisual attention, where effect of co-spatial audio on the salience of a weakly visible target among prominent distractors is studied. We show that co-spatial audio increases the visual saliency of the weak visible target.

In Chapter 7, we describe a feed-forward, proto-object based audiovisual saliency model to detect salient visual, auditory and audiovisual events in dynamic natural scenes. The salient locations identified by the model generally agree with our judgment.

Lastly, we discuss some interesting ideas that can be pursued in future in Chapter 8.

Chapter 2

Spectral Anisotropy is a valid local cue of FGO

2.1 Overview

Objects tend to be convex, opaque and textured, the combination of which leads to the appearance of a feature gradient near the occlusion edge. A mathematical analysis of such feature gradients from a differential geometry point of view was done in [13], following which psychophysics experiments of [95] found that human observers use such cues. Motivated by these studies, using Principal Component Analysis, we found [15] that, next to T-junctions, the strongest local feature in natural images predictive of FGO is a feature gradient on the figure side. In the spectral domain, these gradients are characterized by anisotropic distribution of power in high frequency bins. Our method, which uses fast local 1D DFTs to characterize this new local cue, which we call *spectral anisotropy*, is particularly suitable

for machine vision applications, especially figure/ground (FG) labeling of an image.

We show that this measure by itself is surprisingly informative about FG relationships, with a discrimination accuracy exceeding 60% when applied to a single location on the OB. These measures are related to the already mentioned shading patterns at the edges of objects [13], including extremal edges [14, 15]. The results are replicated on two different databases of image patches, where the characteristics of images in each dataset is different. Next, we show that Spectral Anisotropy of close-by locations along the boundary co-varies but is largely independent over larger distances which allows to combine results from different image regions.

In summary, we show SA is a property of the foreground side of the OB, which arises due to surface markings near the OB on the figural object undergoing spatial compression due to perspective projection. This spectral signature is seen irrespective of the photography technique used (focused on the object *vs.* focused at infinity), size of local neighborhood used for analysis or the size of images (See Appendix A).

We provide a context for our work in Section 2.2, introduce our basic measures in Section 2.3, define them formally in Section 2.4, and apply them to a large set of natural scenes in Section 2.5. We conclude with a Discussion in Section 2.6.

2.2 Related Work

In this study, we investigate the spectral properties of small patches of natural images extracted along the OBs for the purpose of identifying local cues of FG organization. Spectral properties of natural images have been studied from various perspectives. Sev-

eral authors [96, 97, 98, 99] have analyzed the statistics of entire images and shown that the power of the rotationally averaged spectrum varies inversely with spatial frequency, a key property that gives rise to scale-invariance in natural images. van der Schaaf and van Hateren [100] showed that the distribution of spectral power is not isotropic but is higher for horizontal and vertical than other orientations. It was shown that rough depth estimation [101] and limited scene categorization [102] can be performed based on the Fourier energy spectrum of entire images.

For the task of establishing FG relationships that we focus on, spatial frequency as a global cue has been studied behaviorally for more than a half century. Gibson [103] claimed that regions with low spatial frequency are likely to be perceived as figure and those with higher spatial frequency as ground. Contrastingly, in the psychophysical experiments of Klymenko and Weisstein [104], it was found that a region with higher spatial frequency was perceived as figure on more occasions than regions with lower spatial frequency. Note that in these and later behavioral studies, spatial frequency was averaged (separately) over the entire figural and ground regions. These studies did not consider variations of spatial frequency as a function of the distance from the figure/ground boundary nor as a function of orientation, along or orthogonal to the boundary, which is the analysis we perform in the present study. Moreover, in most of these earlier psychophysical experiments artificial stimuli were employed rather than natural scenes.

Fowlkes et al. [11] showed that figural regions are locally smaller and more convex, and that they are often situated below the OB. In a related investigation [105], the convexity of the OB, a local FG cue, was found to increase the perceived depth difference between figure

and ground. Ren et al. [106] used local *shapeme* models to perform FG assignment in natural images. These authors used a logistic classifier algorithm to locally assign figure/ground labels, and a Conditional Random Field based global model to enforce consistency of FG relationships at T-junctions. Their local and global models achieve 64% and 78% accuracy respectively in determining correct FG relationships. Geisler et al. [107] train neurally-inspired models for, among other tasks, FG classification. Different from their work, our approach does not use any training; instead we directly exploit the statistics of natural scenes, as will be discussed below in Sections 2.3 and 2.6. Furthermore, the description of the stimulus encoder "neurons" in [107] is in the spatial domain while we use information in the spectrum. Another difference is that only foliage data is used in the Geisler *et al.* study [107] while we use images of natural scenes from a large number of different scene classes. A more recent abstract also reported results on the interaction of local cues (convexity and closure) for FG segregation [108].

An interesting heuristical approach in the computer vision literature combines various image cues, both local and global, to infer 3D depth information from 2D images [109, 110]. From a set of elementary assumptions, *e.g.,* that neighboring pixels belong to the same surface if there is no edge between them, that long straight lines belong to structures like buildings, sidewalks or windows, that the sky is on top of the image *etc*, Saxena et al. [110] reconstruct a 3D depth map from a single 2D image. Even though these cues are not explicitly designed for FG segregation, inferring the depth-wise arrangement of a scene necessarily leads to the determination of FG relations in many cases.

In summary, previous investigations of FG relationships in the spectral domain have

focused on either the global power spectrum of the entire image, or the local power spectrum in different parts of the image. We decided to extend these approaches by taking into account the location of OBs. In our study, we compute the local spectral power in small patches selected from natural scenes that are adjacent to known OBs, with the goal of determining on which side of the boundary the figure is situated. Specifically, we study the variation of spectral power in different orientations with respect to the FG boundary, parallel and orthogonal to it. Based on our analysis, we devise a simple FG discrimination rule.

2.3 Spectral Anisotropy Close to Object Boundaries

The study of spectral anisotropy (SA) at occlusion boundaries is motivated by the observation of fundamental properties of the physical world. Objects tend to be convex, opaque and textured, the combination of which leads to the appearance of a feature gradient near the OB. As a consequence, there are systematic differences in the local statistics between the areas of the figure and of the ground which are adjacent to the OB. While visual patterns in the background are not affected by the occlusion, features change in a characteristic way on the figure side. Following theoretical work by Huggins et al. [13], Palmer and Ghose [14] showed that the characteristic feature gradients on the figure side (the so-called *extremal edges*) can be used by human observers for figure ground segregation. The components of extremal edges in natural scenes can be identified and classified using Principal Component Analysis [13, 15]. We therefore decided to exploit the predicted differences between feature

gradients along the OB and orthogonal to it by characterizing them in terms of local discrete Fourier transforms and then quantifying localized spectral image statistics on the two sides of the boundary.

We select pairs of image patches of size $K_s \times K_s$ that straddle the OB at a number of locations along the OB. At a given location on the OB, a pair of patches, one located on the figure side and its counterpart on the background side is extracted, see A.1 for the procedure. The pixels on the OB between them are not considered part of either patch. A patch is denoted by $\psi_s(x, y)$, where the subscript s denotes the side of OB containing figure (f) or ground (g),

$$s := \begin{cases} f & \text{if} \quad \psi_s(x,y) \quad \text{is on the figure side} \\ g & \text{if} \quad \psi_s(x,y) \quad \text{is on the ground side} \end{cases} \tag{2.1}$$

Let us define a local coordinate frame in the patch $\psi_s(x, y)$ with x varying parallel to the OB, and y orthogonal to it. The oriented power spectrum parallel to the OB of a patch on side s is defined as,

$$E_{s\|}(u, y) = |\Psi_s(u, y)|^2 \tag{2.2}$$

where $\Psi_s(u, y)$ is the (windowed, see next paragraph) one-dimensional Discrete Fourier Transform (DFT) of $\psi_s(x, y)$ with respect to x (parallel to the boundary, denoted by the symbol $\|$) at distance y from the boundary, and u is the spatial frequency variable corresponding to parallel orientation.

The definition of $\Psi_s(u, y)$ in eq 2.2 is $\Psi_s(u, y) = \mathcal{F}_x \{\psi_s(x, y) \times h(x, y)\}$, where $\mathcal{F}_x(.)$ denotes the 1-D DFT with respect to x, and $h(x, y) = 0.54 - 0.46 \cos(2\pi \frac{x}{K_s})$ is the 1-D

Hamming window applied to row y. Note the absence of any dependence on y in the Hamming window, it is applied to each row independently before computing the DFT to reduce boundary artifacts [111]. Results do not depend critically on the windowing function, we repeated the analysis using a Bartlett window and obtained very similar results.

The average oriented power spectrum of the patch $\psi_s(x, y)$ parallel to the OB is obtained as

$$\overline{E}_{s\|}(u) = \frac{1}{K_s} \sum_{y=0}^{K_s-1} E_{s\|}(u, y) \tag{2.3}$$

The average oriented power spectrum of a patch orthogonal to the OB, $\overline{E}_{s\perp}$, is computed analogously, with the one-dimensional Fourier transform now performed on the y coordinates, and the Hamming window applied correspondingly.

The total oriented spectral power (a scalar) of $\psi_s(x, y)$ in the frequency range $\{u_1, \ldots, u_2\}$, parallel to the OB is,

$$[T_{s\|}]_{u_1}^{u_2} = \int_{u_1}^{u_2} \overline{E}_{s\|}(u) du \tag{2.4}$$

The total oriented spectral power of a patch orthogonal to the OB, $[T_{s\perp}]_{v_1}^{v_2}$, is computed analogously.

The ratio of orthogonal to parallel total oriented spectral power for patch ψ_s, $s \in \{f, g\}$ is defined as the SA,

$$\rho_s(u_1, u_2, v_1, v_2) = \frac{[T_{s\perp}]_{v_1}^{v_2}}{[T_{s\|}]_{u_1}^{u_2}} \tag{2.5}$$

When $\rho_s(u_1, u_2, v_1, v_2)$ is equal to unity, the patch is said to be spectrally isotropic, otherwise it is spectrally anisotropic.

The *unoriented* total spectral power $\bar{T}_s(u_1, u_2, v_1, v_2)$ of a patch is defined as the average

of the oriented spectral powers, $[T_{s\parallel}]_{u_1}^{u_2}$ and $[T_{s\perp}]_{v_1}^{v_2}$. For example,

$$\bar{T}_f(u_1, u_2, v_1, v_2) = \frac{1}{2}([T_{f\parallel}]_{u_1}^{u_2} + [T_{f\perp}]_{v_1}^{v_2})$$

is the unoriented total spectral power of the figure side.

2.4 Data and Methods

We use two image databases freely available on the internet, the MIT LabelMe [112] collection and the Berkeley Segmentation Data Set, BSDS300 [113], to prepare our datasets of image patches.

The BSDS300 database consists of 300 images, all with an image size of 481×321 pixels. The database also contains human-drawn contours along the object boundaries to segment objects in the scenes, with one boundary map per observer and image. Most images have multiple boundary maps. The number of segmented regions varies both across images, for the same observer, and across observers, for the same image. For each image, we chose the boundary map that had the smallest number of segmented parts (five images in the database did not have any associated boundary maps and were not used). The location along the OB (yellow dot in Figure 2.1A) at which $K_s \times K_s$ figure and ground patches were extracted is generated by randomly drawing (without replacement) one location (*i.e.*, one pixel) from among all locations (pixels) in the boundary map. Patches were then rotated to a common orientation such that the orientations orthogonal and parallel to the OB in the image coincide with y- and x-axes respectively of the rotated patch as described in A.1. All

patch rotations were done in the image plane and a bi-linear interpolation scheme [114] was used to compute pixel values at the rotated locations. We collected 5 figure patches and their background counterparts per image, a total of 1475 FG patch pairs from the BSDS300 dataset. We are interested in systematic effects along an OB but not in the influence of structural cues like L-junctions or T-junctions. Therefore, if any of the patches contained a clear T-junction or L-junction, it was replaced by another patch randomly selected from the same contour in the same image. This was the case in 113 out of the 1475 total (81 T-junctions and 32 L-junctions).

The MIT LabelMe database consists of a very large number of user-contributed images with user-labeled objects but without accurate boundary maps. Our goal was to generate a set of images that is representative of a broad range of natural scenes, to avoid a systematic preselection for specific types of patches and to reduce the effect of biases such as illumination, frequently occurring foreground and background types, local curvatures, textures, color variations *etc*. Therefore we selected 585 images from five categories: office environment, other indoor scene (living room, kitchen *etc*), street, beach, and forest. Due to the heterogeneous nature of the database, the selected images varied in size from 256×256 to 2048×1500 pixels. Since no object boundary maps were provided, patch locations were selected on perceived boundaries by a human observer (the first author). Patches were then rotated to a common orientation, see A.1. Again, patches with T and L-junctions were avoided during the selection process. A total of 1761 figure patches and their ground counterparts were collected, with a varying number of patch pairs from each image. Numbers of images and FG patch pairs in the different categories are given in Table 2.1.

Category	Number of Images	Number of patch pairs
Indoor	199	524
Beach	138	480
Office	62	204
Street	120	340
Forest	64	213
Total	**585**	**1761**

Table 2.1: Number of images and figure-ground pairs used from the LabelMe Dataset, by image category.

The patch extraction process is illustrated in Figure 2.1; for the detailed procedure see A.1. The blue and red boxes in Figure. 2.1A enclose a pair of figure and ground image patches respectively in their original orientation. The blue arrow is positioned at the OB location centered on the extracted patches, and it is oriented orthogonal to the OB pointing toward the background. The pair of rotated, bi-linearly interpolated patches, each denoted by $\psi_s(x, y)$ are shown in Figure 2.1B (note pixelation). After rotation, patches are converted from RGB colorspace to 8-bit grayscale where the intensity I of the patch is obtained from the Red, Green and Blue color channels R, G and B as $I = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B$ [115]. All analyses described in Section 2.3 are performed on these grayscale patches.

For all analyses throughout the paper, figure and ground patches of 16×16 pixels are used (other sizes are discussed in A.4), an example is shown in Figure 2.1B. One dimensional DFTs are computed on the figure and ground patches separately, as described in Section 2.3. We compare the distribution of spectral power in a patch between orthogonal and parallel orientations on a one-to-one basis (Equations 2.3 and 2.4). In the Fourier domain, this gives us 8 bins for each orientation. In Sections 2.5.1 and 2.5.2, comparison of both *oriented* and

unoriented spectral power distributions is made between figure and ground.

We study the spectral properties of image patches in the BSDS300 and LabelMe databases separately because of the following reasons: (1) In the set derived from the LabelMe database, the images were contributed by users unknown to database providers, but the location of patches on the OB was hand-selected by a human observer (because boundary maps were unavailable), rather than randomly placed on the boundary. We want to make sure that any biases that may have been introduced by the manual selection of patch locations on the boundaries can be isolated by comparing with the BSDS300 (random selection on the boundaries) results, see Section 2.6 for related discussion. It should be pointed out though that at the time of patch collection (for both datasets), our goal was to identify *all* potential local cues of FG organization and that the human observer was unaware of the potential importance (and even existence) of any SA cues. (2) There are some major differences between the BSDS300 and LabelMe databases. While LabelMe consists of user contributed images of varying complexity and quality, ranging from shots taken by untrained observers with simple point-and-shoot cameras to images composed by professional photographers using high-performance equipment, BSDS300 images are hand-selected by database providers, have rich texture and uniformly smaller (481×321 pixels ≈ 0.15 megapixels) than more than half of LabelMe images (see Table A.2 for LabelMe image sizes). We want to verify SA is not affected by biases that may have been introduced at the time of image selection, see Section 2.6 for related discussion. (3) A wide range of image sizes available in LabelMe and a fixed image size of BSDS300 with human marked boundaries allow us to test the performance of SA as a function of a some relevant parameters. The robustness

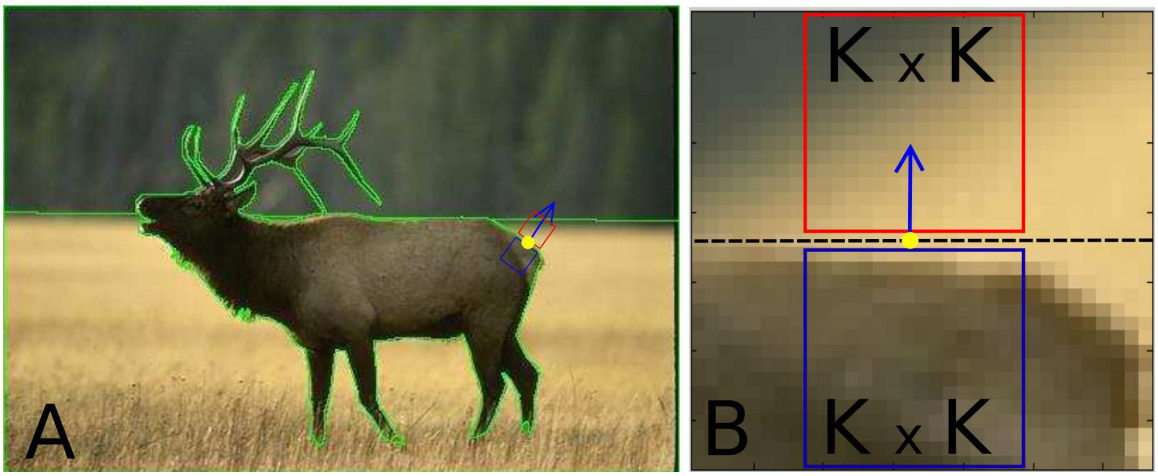

Figure 2.1: Patch extraction. (A) Example image. The green lines are the human-labeled object boundaries, the yellow dot on the boundary is the randomly selected boundary location from which a pair of figure and ground patches is extracted. The blue and red boxes contain figure and ground patches respectively in their original orientation. The blue arrow points towards the background. (B) Image patches after rotation. The boundary of the object is the row of pixels in the center (dashed black line), which is considered a part of neither the figure, nor the ground. The bottom $K_s \times K_s$ blue square is the figure patch (occluding object) and the red top square is the ground patch. A slightly larger area containing figure and ground patches is shown so that context of patch on the boundary is clear. For all analyses except in A.4, $K_s = 16$.

of SA as a FG cue with varying image sizes is discussed in A.5, while its effectiveness as a function of patch size in A.4.

All images and extracted patches labeled for figure and ground are available at `cnslab.mb.jhu.edu`.

2.5 Results

In the following sections, we perform a series of related analyses, using the same statistical procedure in all cases. Using a χ^2 goodness of fit test, we found that distributions are not normal. We therefore used Wilcoxon signed-rank tests, the analogue of paired sample

student's t-tests for normal distributions, in all the following analyses, always with a significance level $\alpha = 0.05$. The number of samples (FG patch pairs) is 1475 for BSDS300 and 1761 for LabelMe databases respectively. For the χ^2 goodness of fit test, 30 bins are used but results are not shown explicitly.

To orient the reader, we begin with a brief summary of the results. In Sections 2.5.1-2.5.3, we first check if mean pixel intensity (power in bin 1) or total unoriented power ($\bar{T}_s(1,8,1,8)$) of patches of the two sides can predict FG relationships. After verifying that those quantities are not useful, we find statistically significant differences in the spectral power of high frequency bins (specifically, $\bar{T}_f(3,8,3,8) > \bar{T}_g(3,8,3,8)$). We then investigate the origin of this difference by comparing oriented spectral power with two orientations, orthogonal and parallel to the OB. We find that in the background there is no difference between oriented spectral powers, parallel and orthogonal to the boundary (as one might expect) while on the figure side, $[T_{f\perp}]_3^8 > [T_{f\parallel}]_3^8$ (which is a novel and, as we later show, useful observation). This effect is what we call SA. A more detailed explanation of the geometrical structure that likely gives rise to SA is given in Section 2.6.

After establishing SA as a valid cue for FG organization, we investigate in Section 2.5.4 how it varies as a function of the distance between patch locations along the boundary. This is important to evaluate the efficiency with which information from multiple boundary locations can be combined to make reliable FG classification decisions. The method is robust to changes in patch and image sizes, as shown in A.4 and A.5.

2.5.1 Basic spectral properties along the boundary

First we test whether there is a systematic intensity difference between the sides, *i.e.* whether the figure side is consistently brighter than the ground or *vice-versa.* We compare the distributions of unoriented spectral power in the first bin (corresponding to mean pixel intensity or DC level) across all patches using a Wilcoxon signed-rank test to verify whether the distribution medians are statistically different [116]. We cannot reject the null hypothesis that the two distributions are identical with similar medians (BSDS300 ($\bar{T}_f(1,1,1,1)$ *vs.* $\bar{T}_g(1,1,1,1)$): $p = 0.20$; LabelMe: $p = 0.66$). Therefore, intensity cannot be used to determine figure ground organization. Next, we compare the distribution of total unoriented spectral power of all figure patches against those of ground patches (*i.e.* $\bar{T}_f(1,8,1,8)$ *vs.* $\bar{T}_g(1,8,1,8)$). Again, a Wilcoxon signed-rank test shows that the null hypothesis (medians are equal) cannot be rejected (BSDS300: $p = 0.32$; LabelMe: $p = 0.67$).

While there are thus no systematic differences in mean pixel intensity or total power on the two sides, we do observe power differences between $\bar{T}_f(3,8,3,8)$ and $\bar{T}_g(3,8,3,8)$. Figure 2.2 shows that the unoriented spectral power (dashed blue and black lines indicating figure and ground respectively) in bins $\{3,\ldots,8\}$ on the figure side is higher than on the background side. Statistical significance tests confirm that $\bar{T}_f(3,8,3,8)$ in figure is greater than $\bar{T}_g(3,8,3,8)$ in ground (Wilcoxon signed-rank tests - BSDS300: $p = 2.79 \times 10^{-24}$; LabelMe: $p = 1.08 \times 10^{-4}$).

A possible explanation for the occurrence of elevated power levels in bins $\{3,\ldots,8\}$ on the figure side is the presence of anisotropic spectral power distributions. Motivated by our and others' observations of differences in the spatial structure on the two sides of an OB

[13, 14, 15], we decided to consider *oriented* spectral power with respect to the OB.

2.5.2 Spectral Anisotropy

We quantify SA in figure and ground separately in two orthogonal orientations with respect to the OB, as detailed in Section 2.3. In Figure 2.2, we show the mean spectra of all patches in the BSDS300 dataset, for an analogous plot for LabelMe see A.2. The figure shows: the oriented power spectra of (1) figure orthogonal to the OB, $\overline{E}_{f\perp}$ (solid green line); (2) figure parallel to the OB, $\overline{E}_{f\parallel}$ (dashed green line); (3) ground orthogonal to the OB, $\overline{E}_{g\perp}$ (solid red line); (4) ground parallel to the OB, $\overline{E}_{g\parallel}$ (dashed red line); and also the unoriented power spectra (dashed blue and black lines representing figure and ground sides respectively). The error bars indicate standard error. We see that for bins 1-2, there are no differences between any of the oriented spectral power levels. Even at higher frequencies (bins 3-8), the mean spectra for the background in both orientations overlap with each other. However, at these higher frequencies, on the figure side, power orthogonal to the OB is higher than parallel to the boundary.

We therefore proceed to compare the oriented power on the two sides only for the high-frequency bins. The distribution of $[T_{s\perp}]_3^8$ *vs.* $[T_{s\parallel}]_3^8$ for all 1475 patches from the BSDS300 data set is shown in Figure 2.3, using blue dots for the foreground ($[T_{f\perp}]_3^8$ against $[T_{f\parallel}]_3^8$) and red dots for the background ($[T_{g\perp}]_3^8$ against $[T_{g\parallel}]_3^8$). The abscissa and ordinate thus represent total power (bins 3–8) parallel and orthogonal to the OB, respectively. The marginals along the two axes seem to show a shift towards higher frequencies of the figure *vs.* the ground both parallel and orthogonal to the edge. The origin of this shift is unclear; it could be due

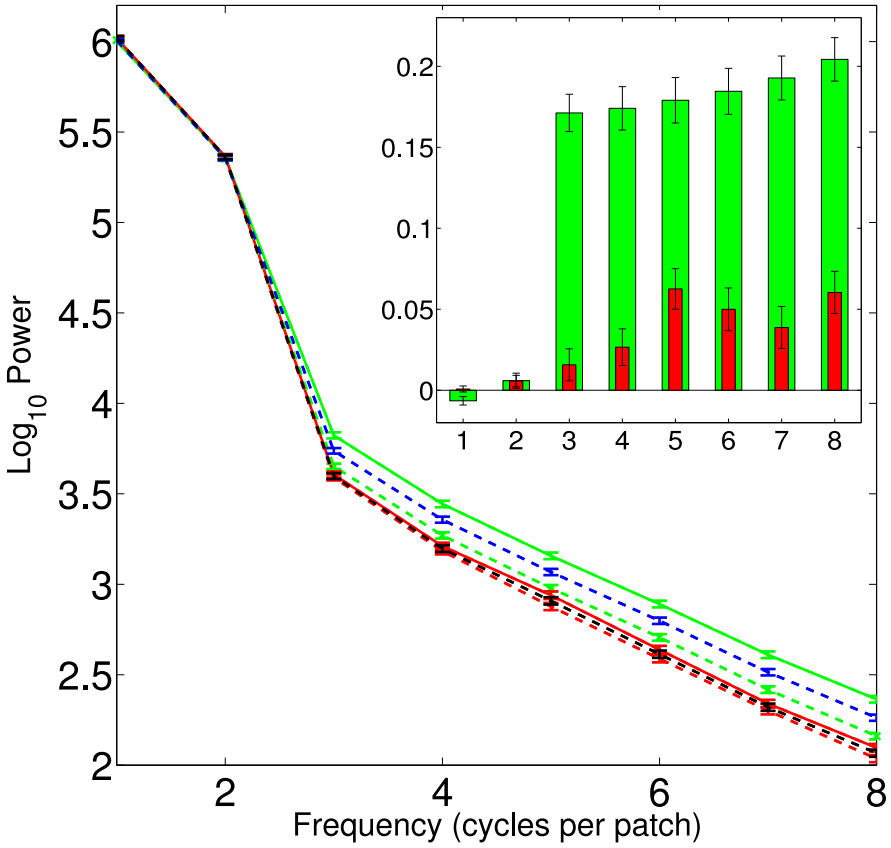

Figure 2.2: Average power spectra of all patches of BSDS300 data as function of spatial frequency. The unoriented spectra are represented by dashed blue (figure) and black (ground) lines. The oriented spectra in the plot are: $\overline{E}_{f\perp}$ (solid green line), $\overline{E}_{f\parallel}$ (dashed green line), $\overline{E}_{g\perp}$ (solid red line) and $\overline{E}_{g\parallel}$ (dashed red line). Inset: The difference in power orthogonal and parallel to the OB ($\log_{10}(\bar{E}_{s\perp} - \bar{E}_{s\parallel})$) as function of spatial frequency. Axes are the same as in the main figure. Green and red bars represent figure ($s = f$) and ground ($s = g$) differences respectively. Error bars are standard errors in figure and inset. Significant differences are only observed for higher frequencies (bins 3-8), and they are significantly larger for the figure than for the ground side. Results from the LabelMe database are similar, see A.2.

to the photographers focusing on the foreground rather than the background, resulting in more power in the higher spatial frequencies on the foreground than the background side. Therefore, we do *not* exploit this effect (which is absent in the LabelMe data, see A.2) for FG segregation. Instead, we observe a bias indicating $[T_{f\perp}]_3^8 > [T_{f\parallel}]_3^8$ on the figure side (blue) relative to the background (red) in the marginals along the diagonal, as predicted by SA. Note that the large range required use of a logarithmic scale for the ordinate for this marginal (but not for the marginals along the axes) which graphically de-emphasizes the size of the effect.

Next, we test if this difference is statistically significant. The comparison is made for the two sides separately. The distributions were found to be non-normal with χ^2 goodness-of-fit tests. A Wilcoxon signed-rank tests indicate that for the figure, the power orthogonal ($[T_{f\perp}]_3^8$) to the OB is higher than that parallel ($[T_{f\parallel}]_3^8$) to the figure/ground boundary (BSDS300: $p = 3.03 \times 10^{-31}$; LabelMe: $p = 2.18 \times 10^{-85}$). In contrast, for the ground, oriented power levels ($[T_{g\parallel}]_3^8$ and $[T_{g\perp}]_3^8$) are not significantly different (BSDS300: $p = 0.72$; LabelMe: $p = 0.26$). This indicates an anisotropic distribution of high frequency spectral power on the figure, but not the ground side. A linear regression model, with slope as the only parameter (forced to pass through the origin), was fitted to the distributions of the $\log_{10}$-transformed power, $[T_{s\parallel}]_3^8$ and $[T_{s\perp}]_3^8$, for figure and ground separately. The model exhibits different slopes, with non-overlapping confidence intervals. The slopes significantly exceed unity on the figure side but not on the ground side. Results for both data sets are shown in Table 2.2.

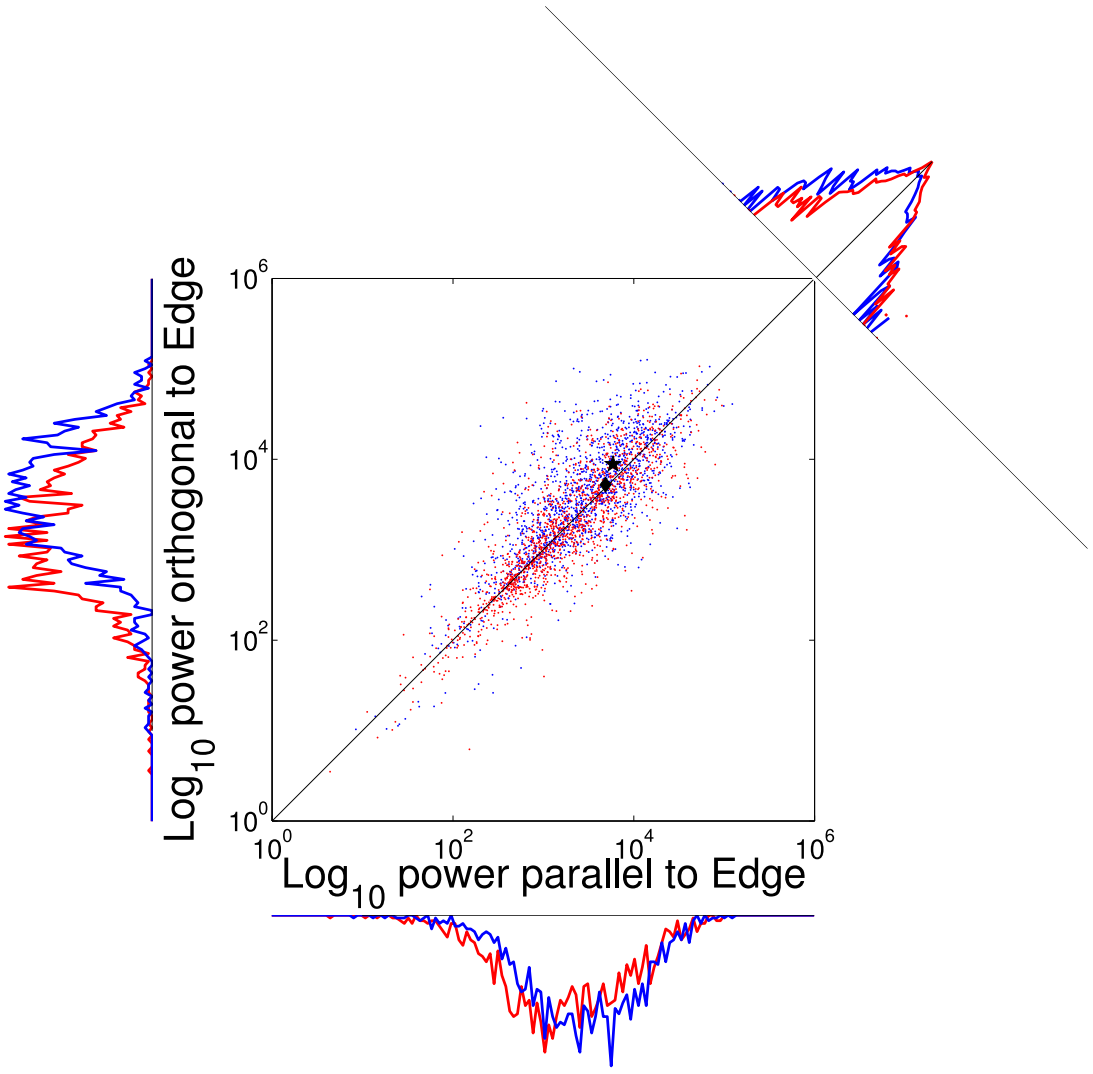

Figure 2.3: Two-dimensional distribution of spectral power in bins 3–8 orthogonal *vs.* parallel to the OB for all BSDS300 patches. Red, background ($[T_{g\perp}]_3^8$ *vs.* $[T_{g\parallel}]_3^8$); blue, figure ($[T_{f\perp}]_3^8$ *vs.* $[T_{f\parallel}]_3^8$). The black diamond, very close to the identity line, shows the mean of the background. The black asterisk, above the identity, shows the mean of the figure. The distance between the figure-side mean and the identity line is even larger for LabelMe, see A.2. The marginal distributions along the scatter plot axes, with linear ordinates, show that average power on the figure side exceeds that on the ground side both parallel and orthogonal to the OB. While this effect seems quite strong here, we do not exploit it for FG segregation since it is absent in the LabelMe data. The marginal distribution at the top right collapses data along the diagonal and has a logarithmic ordinate since the values of the central bins vastly surpass those of other bins. This marginal shows the presence of spectral anisotropy (blue curve above the red one left of diagonal). Again, the effect is stronger in the LabelMe data.

		slope(radians)	CI (low)	CI (high)	R^2
BSDS300	**Figure** (orthogonal *vs.* parallel)	1.036	1.030	1.043	0.53
	Ground (orthogonal *vs.* parallel)	0.998	0.993	1.004	0.71
LabelMe	**Figure** (orthogonal *vs.* parallel)	1.0722	1.065	1.079	0.53
	Ground (orthogonal *vs.* parallel)	1.006	0.995	1.006	0.57

Table 2.2: Regression of $\log_{10}$-transformed high-frequency spectral power in orthogonal and parallel orientations with slope as the only parameter. Results for both datasets show slopes close to unity in the background and greater than unity (and higher than background) in the figure, with their confidence intervals (CIs) non-overlapping. This indicates higher oriented spectral power orthogonal to the boundary than parallel to it on the figure side.

2.5.3 Figure-ground classification based on SA

Can the observed SA be used for determining figure-ground segregation? To answer this question, we developed a FG classification test based on the ratio of oriented spectral powers, bins 3–8. Note that our method does not involve any training, instead, the test is developed from first principles, *i.e.* the statistics of natural scenes discussed above, see also Huggins et al. [13], Palmer and Ghose [14], and then validated on two different data sets.

Let us denote the two sides of a given patch pair by s_1 and s_2 respectively, where s_1 and s_2 can be either figure or ground. Let $\rho_{s_1}(3,8,3,8)$ and $\rho_{s_2}(3,8,3,8)$ be the corresponding ratios (defined in Eq 2.5) of the two sides. We decide whether side s_1 is figure or ground based on the following rule:

$$s_1 := \begin{cases} \text{figure} & \text{if} \quad \rho_{s_1} > \rho_{s_2} \\ \text{ground} & \text{if} \quad \rho_{s_2} \geq \rho_{s_1} \end{cases} \tag{2.6}$$

where we omitted the arguments of ρ_{s_1} and ρ_{s_2}. The classification rule in Equation. 2.6 is a *maximum likelihood* classification rule, where a patch is classified as belonging to one of

the two classes only if its likelihood of belonging to that class is maximum (see A.3 for a detailed explanation). The test yields a classification accuracy of 62.57% for the BSDS300 and 64.51% for the LabelMe datasets respectively. This is a central result of our study. Inverting the pixel intensities on both sides gives very similar results (BSDS300: 61.15%, LabelMe: 66.21%). This again indicates that SA is purely a function of spatial frequency content of the local patch along the OB and that mean pixel intensities have no influence on the properties observed.

As an illustration of FG classification results, a sample of 8 images from the BSDS300 database is shown in Figure 2.4. Half of the images show a sharp background (large depth of field, DOF) and the other half a blurry background (small DOF). A blue rectangle is drawn around the patches; green and red arrows indicate correct (pointing to figure side) and incorrect (pointing to ground side) classifications, respectively, based on SA (Section 2.5.2 and Eq. 2.6). The length of an arrow is proportional to the ratio $\rho_{s_1}(3,8,3,8)/\rho_{s_2}(3,8,3,8)$ and signifies confidence in the decision. Figure-ground classification is effective both in images with small and with large DOF.

2.5.4 Combining multiple classification decisions

Can evidence about FG relations from multiple patches along an OB be combined to improve the reliability of the classification? The extent to which this is possible is determined by the degree of dependence between decisions at individual patches. Here we study the simplest case of pairwise correlations between figure/ground classification decisions at two locations that are R_{ob} pixels distant on the OB.

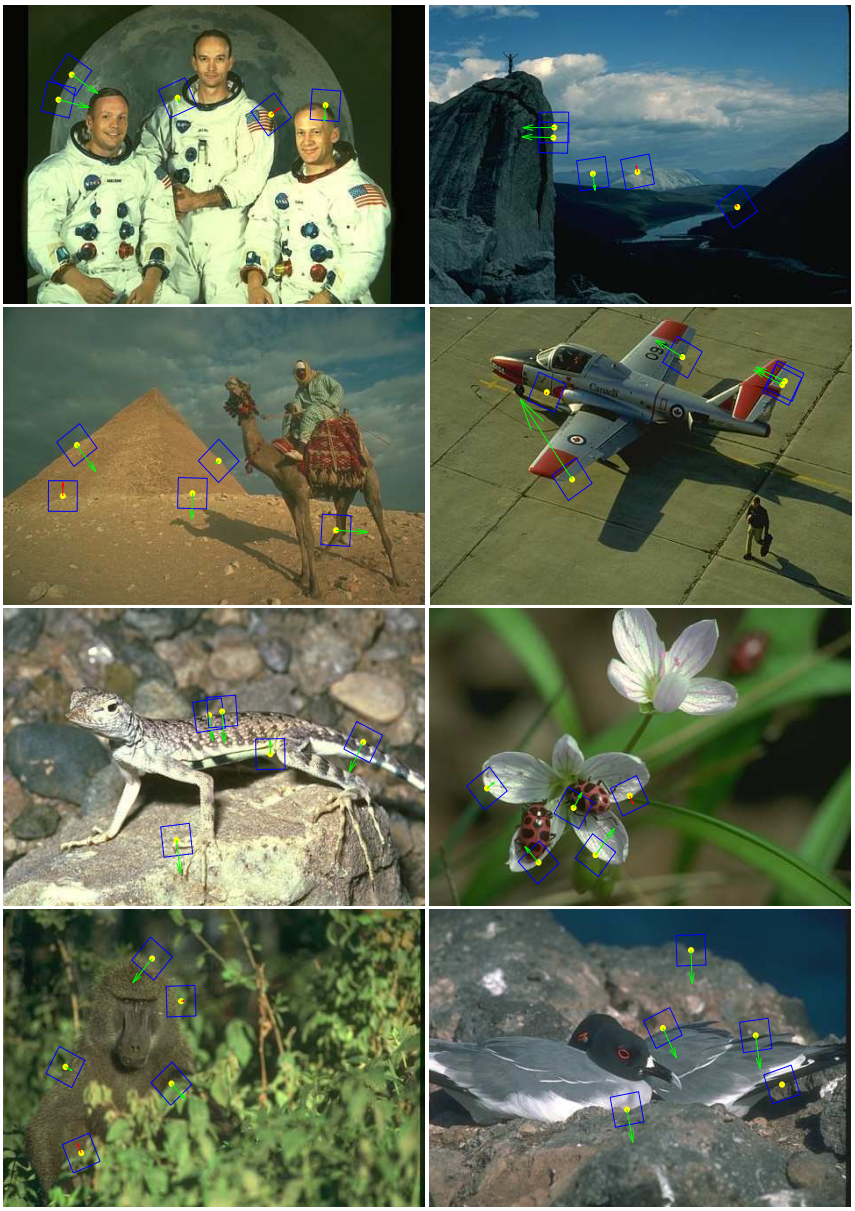

Figure 2.4: Example FG assignments for BSDS300. Blue boxes indicate figure and ground patches, green arrows correct FG assignment, red arrows incorrect assignment. Length of the arrows indicates the confidence level in our classification method. The top four images have large DOF where the entire scene is in focus. In the bottom four images a specific object is focused by the photographer leaving the rest of the scene blurred. As the arrows on the randomly selected patches show (most are green), spectral anisotropy is a useful indicator of figure *vs.* ground.

The analysis is done for the BSDS300 database only. We used the same 1475 figure/ground patch pairs from Section 2.5.1 which we call Dataset 1. In an image, $I^j(x, y)$, $j \in \{1, \ldots, 300\}$, at location $\mathbf{u}_i^j = (x_i^j, y_i^j)$ (yellow dot in Figure 2.5) on an OB in Dataset 1, we draw a circle (dashed black) whose radius R_{ob} is a random number uniformly distributed between 1 and 50. The circle intersects the boundary in, at least, two points. One of these is selected randomly (with equal probability) and the figure/ground patch pair at this location $\mathbf{u}_k^j = (x_k^j, y_k^j)$ (red dot in Figure 2.5) is an entry in a new set of 1475 image patch pairs that we call Dataset[1] 2.

Let d_i^j be a figure/ground classification decision associated with a pair of figure/ground patches at $\mathbf{u}_i^j$, where $d_i^j = 0$ stands for a correct and $d_i^j = 1$ for an incorrect decision. The expectation value $E[d_i^j]$ of d_i^j is the probability of classification error, $P_e(d_i^j)$. For the BSDS300 data set, the probability of correct classification is 0.625, therefore the probability of error $P_e(d_i^j)$ is 0.375. If d_i^j and d_k^j are independent, the joint probability of error is $P_e(d_i^j, d_k^j) = P_e(d_i^j)P_e(d_k^j)$. But, if they are not independent, by the definition of covariance, $P_e(d_i^j, d_k^j) = E[d_i^j]E[d_k^j] + \sigma_e(d_i^j, d_k^j)$, where $\sigma_e(d_i^j, d_k^j)$ is the co variance between d_i^j and d_k^j. We therefore determined the error co variance, $\sigma_e(d_i^j, d_k^j)$ between all possible decision pairs (d_i^j, d_k^j) within each image to see at which distance R_{ob} between patches the error co variance term is small enough so we can drop it.

A plot of $\sigma_e(d_i^j, d_k^j)$ *vs.* R_{ob} is shown in Figure 2.6. The covariance is positive for small distances (as could be expected), and then falls off quite rapidly. For $R_{ob} \approx 30$, at which the

[1] A technical note on patch selection: Although all patches within an image from Dataset 2 are used with all patches in the same image from Dataset 1 and therefore the distance between two patch locations can be as small as one pixel and as large as the largest distance in the image, we select Dataset 2 locations within 50 pixels from Dataset 1 locations to increase the number of close patch pairs. This bias allowed to obtain sufficient numbers of samples for distances up to about 200 pixels to obtain meaningful results. For distances exceeding 200 pixels only few patch pairs were found and they were not included in the analysis.

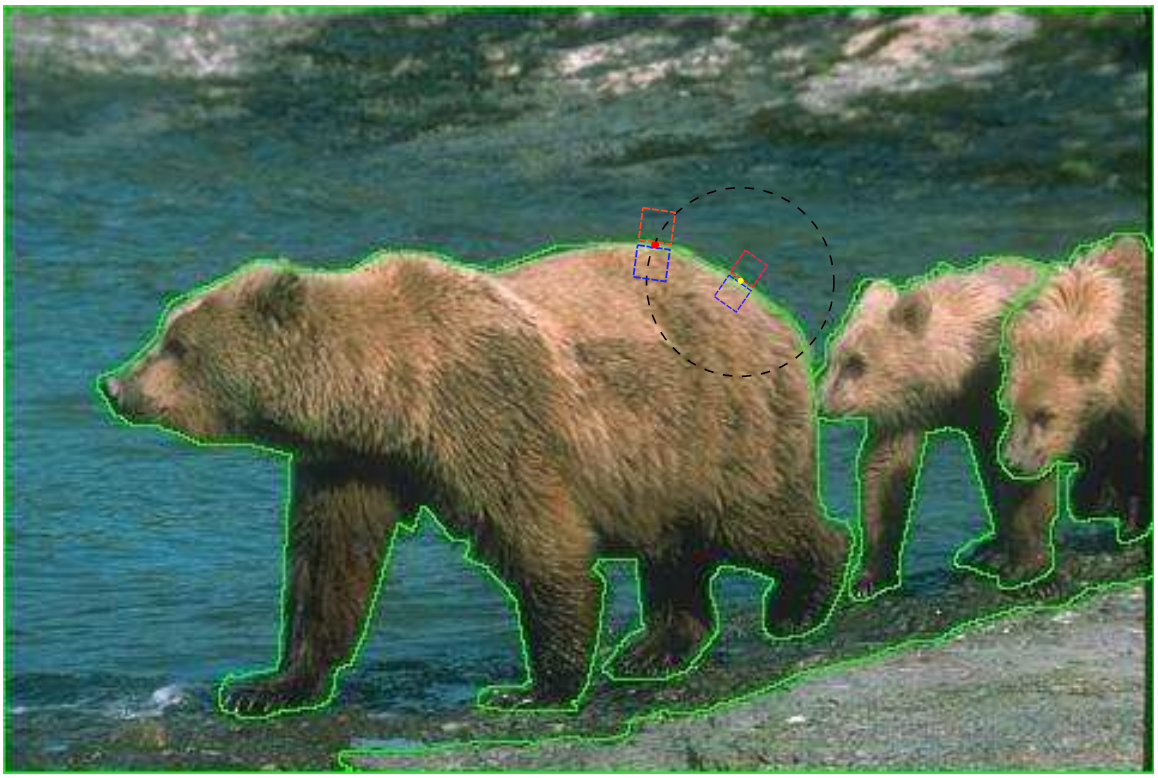

Figure 2.5: Extracting Dataset 2 patches. The yellow dot on the boundary (green line) is the center of one patch pair (red and blue squares) from Dataset 1. A circle of radius R_{ob} (black dashed line) is drawn around it, and one of its intersections with the contour (red dot) is selected as the center of a new patch pair which is entered into Dataset 2.

two locations on the boundary are separated by a distance double the patch size, covariance is already quite small, about 0.02. Values remain mainly positive until $R_{ob} \approx 100$ which may reflect the average size of objects in the images. Beyond this distance, correlations fluctuate around zero.

In conclusion, covariance analysis shows that decisions are only weakly correlated at distances exceeding about twice that of a patch ($R_{ob} \gtrsim 2K_s$). Such patches can be regarded as independent, and the FG decisions from those locations can be combined to improve classification reliability.

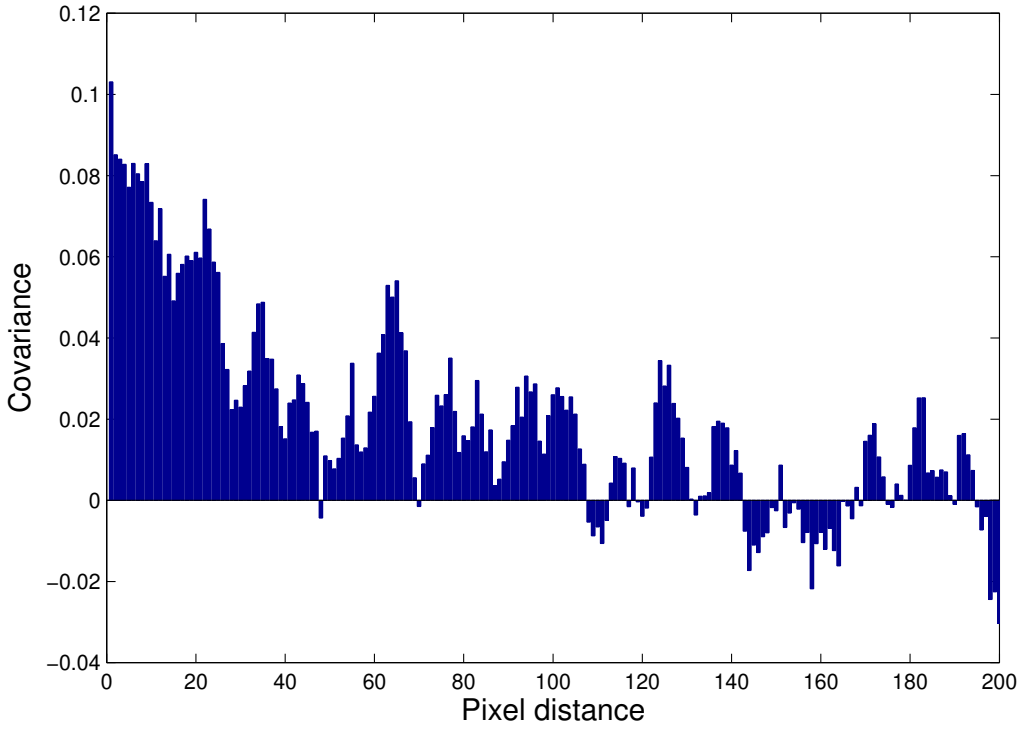

Figure 2.6: covariance of classification decision, $\sigma_e(d_i^j, d_k^j)$ *vs.* pixel distance R_{ob} along the OB . The covariance is high and positive for small distance, less than 5 pixels. Covariance drops off as R_{ob} increases. For distances up to about 100 pixels, $\sigma_e(d_i^j, d_k^j)$ is mainly positive after which it is small and fluctuates randomly. A smoothing running average filter of width 7 is used.

2.6 Discussion

Our results show that spatial frequency power on the figure side perpendicular to the object boundary exceeds that parallel to the object boundary at high spatial frequencies, and that no such difference exists on the background side. We believe that this difference is caused by the visual compression of features on the figure due to the surface curvature of convex objects, as discussed in the next paragraph. Therefore, this SA measure can be used to distinguish the figure from the ground side. We also show that no statistically significant difference exists for the lower spatial frequencies, including mean intensity.

The physical background for our observation is, we believe, the simple fact that most objects are convex. At the object boundary, surface markings on the object undergo spatial compression due to perspective projection. As a result, uniform features spread across a larger region in depth get packed into smaller viewing angles on the figure due to the surface curvature. This means, within the same viewing angle, more high frequency content is present on the figure side compared to the ground side. This only occurs in the orientation orthogonal to the boundary but not parallel to it, resulting in the observed anisotropy. This feature of the visual world has been observed previously in the spatial domain [14] and shown to be useful for FG segregation [13, 15] but ours is the first study to make use of it in the spectral domain. This is important since the effect is straightforward to quantify in spectral terms, as we show in this report. Furthermore, the computation is made very efficient by the use of Fourier techniques, and thus suitable for machine vision applications.

Given that all image data we used are taken in one way or the other by a human photographer who is likely to have controlled, among other parameters, the depth of field,

an important concern is that he or she may have selected to focus on the foreground object while leaving the background blurry [117]. An algorithm for FG segregation that relies on this difference would be of limited use since it would be aided by the photographer's decision. It is therefore of importance to make sure that our algorithm does not rely on this cue. We confirmed that this is the case by three different, independent analyses.

First, we observe that differences in focus between figure and background can explain differences between the spectral powers of the figure and the ground side, the quantity we use is anisotropy on the figure side only. While a photographer may treat figure and background differently, he or she cannot control the oriented spectra (orthogonal and parallel to the OB) separately on either figure or ground since the orientation of the OB varies with each foreground object (and, of course, with each patch in each object).

Second, we are not looking for anisotropy along any arbitrary orientations on figure or ground, but along a specific set of orientations, chosen *a priori*. Our decision to compare spectral powers in orthogonal and parallel orientations in relation to the OB is based on theoretical considerations about statistics of the physical world [13] and on empirical psychophysics [14]. The pattern of feature gradients at the OB expected from these results will give rise to maximal differences in power between directions parallel and orthogonal to the OB, not in some other arbitrary set of orientations. This can be seen directly in the 2D mean power spectra of foreground and background patches analyzed separately. We find that the spectral power orthogonal to the OB substantially exceeds that parallel to the OB on the figure side, while there is no such difference on the background side. Results are shown in A.8, specifically Figure A.7. The effect is clear for both BSDS and LabelMe.

Third, we generated a subset of patch pairs by removing those in which either the foreground or the background was rendered blurry. This yielded a set of 1025 patch pairs for BSDS (out of the 1475 total) and 1716 for LabelMe (out of 1761). We then re-performed the analysis described on the remaining sharply focused patches. We essentially replicated the results obtained for the full set of patches, results are shown in A.6.

Together, we can conclude that the observed SA cannot be an artifact of a particular photography technique.

Another possible confound that needs to be addressed is the effect of the rotation of the image patches which is necessary to perform the Fourier transforms efficiently. Patch rotation by arbitrary angles, as is necessary due to the arbitrary orientations of the OBs, results in pixels being placed in "non-integer" locations relative to the grid defined by the image. When re-aligning the pixels, their values need to be interpolated. The simplest method is to replace pixel values by that of their nearest neighbors. We found that this leads to excessively jagged patches. We therefore used a simple bi-linear interpolation scheme [114] to determine the rotated pixel values. Is it possible that rotation followed by bi-linear interpolation creates a bias in the statistics? To answer this question, let us consider the effect of each operation on an isotropic field of pixels. Since rotation is a rigid transform, no bias with respect to rotation angles is introduced, so whatever was isotropic before remains isotropic after rotation. In bi-linear interpolation, the weights used for calculating the pixel value at the rotated position from its four neighbors depend on the rotation angle. But all pixels in figure and ground patches are rotated by the same angle, therefore the weights will be the same for all "new" rotated pixel locations. Hence,

an isotropic field remains isotropic after bi-linear interpolation. So, a rotation followed by bi-linear interpolation transforms an isotropic field of pixels into another isotropic field, hence no bias/anisotropy is introduced by these set of operations. We also note that bi-linear interpolation has a low-pass filtering effect, as is the case to some extent with other interpolation schemes [118, 119]. But the low-pass effect in a given patch pair will be the same for both figure and ground, since both are rotated by the same amount. As our signal is the difference in high-frequency oriented powers between figure and background side, and since both sides are treated equally in the rotation process,no systematic bias is introduced.

As mentioned in the last paragraph of Section 2.4, one reason for our decision to analyze two different image databases (BSDS300 and LabelMe) separately was to verify that no unintentional biases were introduced in LabelMe by the human observer (S.R.) who selected patches along the object boundaries. The overall agreement in results from the two databases indicates that this is, indeed, the case. Another reason was to verify SA is not influenced by any potential bias in the type, size or quality of images. Again, consistency of results between data sets indicates that this is not the case. However, the effect of SA is less pronounced in BSDS300 than in LabelMe. The small size of BSDS300 images (≈ 0.15 megapixels) could be a possible reason, since more global information is included in the patch when image dimensions are small. The FG classification accuracy of BSDS300 images (62.5%) and that of the subset of LabelMe images which are of comparable size (< 0.5 megapixels) are very close (classification accuracy: 63.3%), further strengthening the argument.

The FG classification accuracy we obtain compares favorably with other stand-alone lo-

cal cues. As we are not aware of any previous work where spectral properties of local regions on both sides of the boundary were used to make FG classifications, a direct comparison with previous work is not possible. However, the best FG classification accuracies reported in [11] for local cues such as convexity (60.1%), size (64.4%) and lower region (67.8%) are in the same range as ours. Furthermore, the method used in that study required the training of a logistic classifier model, whereas our's requires no training. Proper training has the potential to obtain better classification results, see Chapter 3 where we show a non-linear Support Vector Machine based classifier model trained on the same data improves the FG classification accuracy substantially. But, it is worth reminding ourselves, training on realistic data sets usually requires substantial computational effort, much more than methods which can be derived directly from the statistics of natural scenes. Another advantage of methods based directly on hypotheses about natural scene statistics, without intercalation of training procedures, is that they usually allow to draw more direct conclusions about the validity of these hypotheses.

An interesting new observation is the covariance of FG classification decisions along the boundary. Since the classification is based on spectral properties of figure and ground sides, it reveals information about the variation of these properties along the boundary. Spectral properties of neighboring patches are correlated, hence there is some dependence between decisions at neighboring locations along the boundary. Beyond a certain distance (about twice the overlap of neighboring patches, $R_{ob} \gtrsim 2K_s$), the spectral properties become essentially independent. This allows us to combine results from different locations on the same boundary to obtain more accurate results.

We finally address the question whether SA mechanisms may be exploited in biology. Spectral anisotropy captures variations in intensity gradients as well as texture variation. Both of these phenomena have a common cause – the curvature of the underlying surface. It may be possible that neurons are sensitive to such cues, meaning that they are selective to gradients in spatial frequencies. Indeed, responses of neurons in the primate parietal cortex have been reported to correlate with texture gradients compatible with 3D depth perception of tilted surfaces [120]. Responses were invariant over different types of texture patterns and most of these neurons were also sensitive to a disparity gradient, suggesting that they play an important role in the perception of 3D shapes. These or similar neuronal populations may implement the local mechanisms studied in this report and thus complement the global FG segregation mechanisms observed in extrastriate cortex [6, 40, 80]. We explore one such possibility with the spatial frequency selective Complex cells found in area V1, which is covered in detail in Chapter 4.

2.7 Conclusion

An analysis of spectral properties of local image patches in the context of figure ground organization is presented. The oriented high frequency spectral power distribution close to the occlusion boundary is mostly uniform in the background, whereas differences are shown to exist in the figure. For the figure side, the oriented high frequency spectral power orthogonal to the boundary exceeds that parallel to it. The figural spectral anisotropy can thus be used for figure ground discrimination. A statistical test of the ratio of orthogonal to parallel high frequency spectral powers discriminates figure from ground with 60% or

greater accuracy per patch, in both datasets tested. Spectral anisotropy co-varies for close-by locations, but mostly independent over larger distances along the boundary and robust to variation in patch or image sizes.

Chapter 3

Improving Figure-Ground Classification with non-linear SVMs

3.1 Introduction

Spectral Anisotropy, a property observed on the foreground side of the OB is characterized by the spatial frequency power on the figure side perpendicular to the object boundary significantly exceeding that parallel to the object boundary at high spatial frequencies. But, no such difference exists on the background side. Studying the oriented spectral properties on figure and ground sides, designing a classification rule based on SA was the focus of Chapter 2, where we show SA is a valid cue for FGO in natural images. In Section 2.5.3 of Chapter 2, we used a linear discrimination rule based on the SA property, namely the

difference in ratios of spectral power of figure and ground sides. We arrived at this rule based on statistical significance tests and regression analysis. From the logarithmic plots of total power in high frequency bins (Fig 2.3), we see that there is an overlap in the distribution of these ratios in two dimensions. We therefore hypothesized that in a higher dimensional space, the distributions of the four spectral powers may be well separated. Therefore we go from the ratios of spectral powers to a four-dimensional space (the four spectral power levels) and train a Support Vector Machine (SVM) model based on the 4D Spectral Anisotropy features. We use a non-linear classification rule by training the model with radial basis function kernels. Compared to the figure-ground classification accuracy obtained in Chapter 2, which was exceeding 60% per patch, we see that there is a considerable improvement when the non-linear SVM model is used for classification. We use the same set of patches that we used in Chapter 2, and see the improvement in classification accuracy in both the databases of patches is consistently higher than those in Chapter 2.

3.2 Feature Vector Computation

The feature vector is computed by concatenating the total oriented spectral powers computed for figure and ground side patches in orthogonal and parallel directions as explained in Equation 2.4. The four oriented spectral powers considered are, $[T_{f\parallel}]_{u1}^{u2}$, $[T_{f\perp}]_{v1}^{v2}$, $[T_{g\parallel}]_{u1}^{u2}$ and $[T_{g\perp}]_{v1}^{v2}$, where u_1 and u_2 are lower and upper cut-off frequencies in the parallel direction and v_1 and v_2 the cut-off frequencies in the orthogonal directions respectively, as in Chapter 2, Section 2.3.

We now define the 4–dimensional feature vectors used for training the SVM model.

Their elements are the total oriented powers parallel and orthogonal to the figure-ground edge and the feature vectors are thus defined as,

$$\mathbf{f} = \left[[T_{f\parallel}]_{u_1}^{u_2} \qquad [T_{f\perp}]_{v_1}^{v_2} \qquad [T_{g\parallel}]_{u_1}^{u_2} \qquad [T_{g\perp}]_{v_1}^{v_2} \right]^T \tag{3.1}$$

Clearly, the first two dimensions originate from the figure and the last two from the ground. Odd and even numbered indices correspond to the orthogonal and parallel orientations, respectively.

3.3 Support Vector Machines

Support Vector Machines (SVMs) are a supervised binary classification technique [121, 122]. The basic idea is that the classification hyperplane is determined by maximizing its distance from the nearest data points (*support vectors*) on either side of the hyperplane.

Let $\{\mathbf{f}_i : i = 1, ..., n\}$ be the training data set, where $\mathbf{f}_i \in \mathbf{R}^d$. Let the labels on each sample be $y_i \in \{+1, -1\}$, indicating one of the two classes to which the sample belongs. An SVM finds weights, $\mathbf{w}$ and a bias, b, that minimize the Euclidean norm $\|\mathbf{w}\|$ such that, for all data points $(\mathbf{f_i}, y_i)$,

$$y_i(\langle \mathbf{w}, \mathbf{f}_i \rangle + b) \geq 1 \tag{3.2}$$

The support vectors are the feature vectors $\mathbf{f}_i$ that lie exactly on the decision boundary (closest to the classification hyperplane). Hence, for support vectors, we have $y_i(\langle \mathbf{w}, \mathbf{f}_i \rangle + b) = 1$. In the case of nonlinear SVMs, for instance, for SVMs with radial basis function

(RBF) kernels, the inner product is replaced by an appropriate kernel function.

After training, test feature vectors are presented to the SVM. The class a given test vector $\mathbf{z}_j$ is assigned to is computed as $\text{sign}(\langle \mathbf{w}, \mathbf{z}_j \rangle + b)$. More explicitly, $\mathbf{z}_j$ is classified as from class ($y_j = +1$) if,

$$\text{sign}(\langle \mathbf{w}, \mathbf{z}_j \rangle + b) = +1 \tag{3.3}$$

and from class (-1) if this expression is -1.

Finding optimal weights, $\mathbf{w}$ and bias b requires solution of the following optimization problem:

$$\underset{\mathbf{w},b,\xi}{min} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \tag{3.4}$$

$$\text{subject to} \quad y_i(\mathbf{w}^T\phi(\mathbf{f}_i) + b) \geq 1 - \xi_i, \tag{3.5}$$

$$\xi_i \quad \geq 0 \tag{3.6}$$

where the function $\phi(.)$ maps the lower dimensional feature vector $\mathbf{f}_i$ into a higher dimensional space, and ξ_i are slack variables. The bounding box (or penalty term) parameter $C > 0$ determines the accepted rate of misclassification, with lower values leading to more misclassification. Finally, $K(\mathbf{f}_i, \mathbf{f}_j) = \phi(\mathbf{f}_i)^T\phi(\mathbf{f}_j)$ is the kernel function. We use radial basis functions with a nonlinear kernel defined as,

$$K(\mathbf{f}_i, \mathbf{f}_j) = \exp(-\gamma\|\mathbf{f}_i - \mathbf{f}_j)\|^2), \ \gamma > 0 \tag{3.7}$$

where γ is the parameter that controls the width of the kernel. For more detailed discussions of SVM techniques see refs. [123, 124].

3.4 Data and Methods

We used the same MIT LabelME database [125] and the Berkeley Segmentation Data Set (BSDS300) [11] of image patches, which were used in Chapter 2.

As before, our interest is not to find the contour separating figure from ground, a task that we assume has been completed, but to determine which side of the contour corresponds to the figure and which to the ground. Therefore, we trained the SVM model with a set of patches with positive (correct) and negative (inversed) figure-ground assignment. For each patch, the 4-dimensional feature vector $\mathbf{f}$ described in Section 3.2 was computed. We chose $u_1 = 3$ and $u_2 = 8$ in Equation 2.4, similarly $v_1 = 3$ and $v_2 = 8$ for the orthogonal orientation in computing the total oriented spectral powers. Features were centered and scaled to have unit standard deviation. We use radial basis function (RBF) kernels, see eq. 3.7. A subset of vectors, $\mathbf{f}'$, were selected as correctly assigned and inserted into a matrix such that each column corresponds to a feature and each row to a sample. A set, of the same size, of inversely assigned patches was constructed by interchanging the indices corresponding to figure and ground. Training of the SVM then proceeded as described in Sec. 3.3, applying Sequential Minimal Optimization (Matlab, Natick MA). A fraction of 5% of the training examples were allowed to violate the Krush-Kuhn-Tucker conditions [124].

The analysis was carried out for LabelME and BSDS databases separately. The patch databases (1761 patches for LabelME, 1475 patches for BSDS300) were divided into training

and test sets. Two thirds of the patches were used for training and the remaining one third for testing, and selection for these subsets was random. The training patches are further divided into positive examples and negative examples (50:50 ratio) by switching the indices of features as described before. Appropriate class labels, $positive = +1$ and $negative = -1$ were assigned.

The performance of the classifier with RBF kernels depends on two key parameters, γ which determines the width of the RBF kernel, and C which controls the accepted misclassification rate, see eqs. 3.4-3.7. Optimal parameters were determined in a grid search with initial values of both γ and C in the range $[10^{-5}, 10^{5}]$ and initial step size of 1 in the exponent. For each pair of values, we train the classifier with ten fold cross-validation. The trained model with the best cross validation score is used to further refine the values of (γ, C) pairs on a finer grid, where each parameter is systematically varied in small increments (step size 0.1 in the exponent). Once we obtain the optimal parameter values (γ_{opt},C_{opt}), they are used to train the full training set to get our final classifier model.

3.5 Results and Discussion

The optimal values (γ_{opt},C_{opt}) were slightly different for the two databases. They are shown in Table 3.1 together with the respective cross-validation scores, defined as the accuracy of classification obtained on the training set after the 10-fold cross-validation.

Performance of the trained classifiers was assessed by running them on test data. Our test datasets consisted of 491 samples for the BSDS database and 587 samples for the BSDS database, the results are shown in Table 3.1. Compared to the accuracy near $\approx 60\%$ for both

Database	# Samples	γ_{opt}	C_{opt}	CV score	Accuracy
LabelME	587	0.759	1.777	84.07%	67.12%
BSDS	491	0.687	2.282	88.72%	69.25%

Table 3.1: SVM results. For both databases (column 1), we show the number of image patches in the test set (col. 2) and the optimal parameters γ_{opt} (col. 3) and C_{opt} (col. 4). Column 5 shows the cross validation (CV) scores and column 6 the percentage of correct figure-ground assignments.

LabelMe and BSDS datasets that we saw when the simple ratio based classification rule (Eq 2.6) was used in Chapter 2, here with a training-based strategy with SVM classifiers, our accuracy levels increased to 67.12% for LabelMe and 69.25% for BSDS. For the same BSDS database that was used in [11], which we also use in our work, and for the case of a stand-alone local cue, our SVM based FG classification model performs better than all the local cues reported in [11]. This is the highest Figure-Ground classification accuracy for a stand alone local cue reported so far, exceeding other cues such as, convexity, lower region, *etc.* We also ran the classifiers after training on the entire patch databases comprising both training and test data (1761 patches from LabelMe, 1475 from BSDS300). As expected, performance was increased, reaching 74.29% and 73.59% accuracy for BSDS300 and LableMe datasets, respectively.

As noted in Section 3.4, we assume marking of the contour separating the figure and the ground has been completed. We were concerned about possible artifactual results due to slight but systematic misalignment between the actual figure-ground border and the human-generated contour. We therefore applied the already trained SVM model (γ_{opt},C_{opt}) on new data sets generated by shifting the figure and ground parts away from the figure-ground boundary by $1-3$ pixels. We found that this essentially did not change the results, see

Database	1 pixel	2 pixel	3 pixel
LabelME	65.76%	66.61%	67.12%
BSDS	67.21%	68.84%	68.02%

Table 3.2: Accuracy with patches shifted by 1, 2 and 3 pixels (see text).

Table. 3.2.

3.6 Conclusion

We show that the figure-ground classification accuracy can be further improved by moving to a higher dimensional space and employing a non-linear separation hyper-plane. The non-linear Support Vector Machine model with radial basis function kernels improves the classification accuracy remarkably, reaching an average of nearly 70% correct classification per patch, which is consistently found in two different databases with different properties such as image composition, image sizes, *etc.* This is the highest Figure-Ground classification accuracy for a stand alone local cue reported so far, exceeding other cues such as, convexity, lower region, *etc.* This indicates the high frequency oriented spectral powers are excellent features for detecting figure-ground relations in natural images. In addition, we also verify that there are no labeling artifacts as the classification accuracy is maintained nearly the same for different shifts of pixels near the boundary. In future we would like to investigate if there is room for further improvement with the use of more sophisticated machine learning techniques such as random forests, deep learning, *etc.*

Chapter 4

A biologically plausible method of Spectral Anisotropy computation

4.1 Introduction

The computation of SA, as we did in Chapter 2, from an image patch (foreground or background side) involved computing the oriented 1D Discrete Fourier Transform, followed by averaging only the high-frequency spectral powers over the entire patch before computing the ratio of orthogonal to parallel spectral powers. How can this computation be carried out in the visual cortex? Where in the visual system and by what neurons is such a computation carried out? Can such a computation be made possible with only the Simple and Complex cells (see Section 1.1.1) found in area V1? Or does this involve feedback or recurrent connections from higher visual areas?

Previous work has shown that neurons in area V1 are responsive to textures as well

as figure-ground relations [126]. Tsutsui et al. [120] show neurons in the inferior temporal cortex (IT) can detect gradients of texture and code for 3D surface properties. Willmore et al. [44] show that asymmetric RFs in V1 and V2 are selective to figure-ground organization. Sakai et al. [127] demonstrate their model based on asymmetric RFs can detect figure-ground relations. Not all neurons in area V1 fit the description of Simple and Complex cells with a fixed (2 or 3) number of lobes [128, 129]. Many neurons in area V1 have RF properties, different from the end-stopped, hyper-complex cells, Simple or Complex cells found in area V1 [129, 130]. Also, many neurons in the primary visual cortex are sensitive to higher spatial frequencies (more than 5-6 lobes [131]) and second order pixel intensity statistics [28].

In this chapter, we show that a group of V1 Complex cells with overlapping RFs is sufficient to compute SA. A simple method, in which, pooling of the Complex cell responses of same orientation (the filter orientation being parallel to the OB, in order to capture orthogonal spectral power), but different scales, hence spatial frequencies is presented, which can detect SA of the figure side and lead to FGO.

4.2 SA by pooling Complex cell responses

Spectral Anisotropy, a local cue for FGO, that captures intensity and texture gradients very close to object boundaries, is computed by pooling Complex cell responses of various spatial frequencies from small image regions on both sides of the boundary (Figure 4.1).

A Complex cell at location (x, y) maximally responds if an edge of orientation θ is present at (x, y) or if the spatial frequency characteristics of the underlying image region

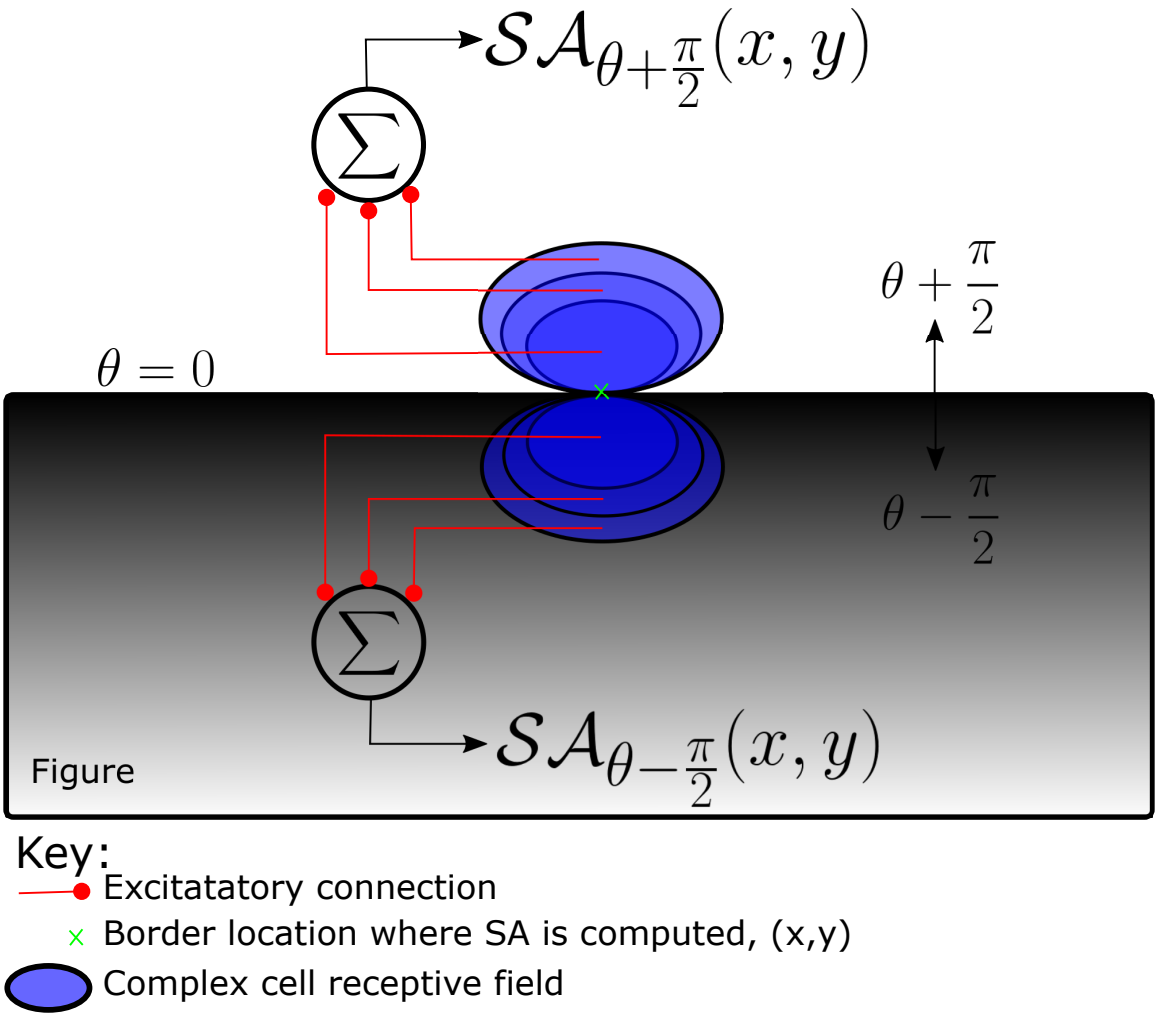

Figure 4.1: Biologically plausible computation of Spectral Anisotropy by pooling Complex cell responses. The local orientation, θ of the border between figure and ground at boundary location (x, y) is 0. SA is computed at (x, y) for two two opposite BO directions, $\theta + \frac{\pi}{2}$ and $\theta - \frac{\pi}{2}$. There is a vertical intensity gradient on the figure side along the horizontal edge. By pooling the complex cell responses at various scales (hence, different spatial frequencies) on the side of an edge, we can quantify the intensity and texture gradients in the direction orthogonal to the edge orientation

matches the RF properties of the Complex cell. Two Border Ownership (BO) directions are possible at (x, y). So, one BO direction will be, $\theta + \frac{\pi}{2}$ (a 90° counter-clockwise rotation with respect to θ), and the other is $\theta - \frac{\pi}{2}$ (a 90° clockwise rotation with respect to θ). We use the same convention throughout this chapter, as well as in Chapter 5. SA at any location, (x, y) in the image, for a specific orientation, θ and for the BO direction, $\theta + \frac{\pi}{2}$ is computed for one side of the border (side determined by the BO direction, $\theta + \frac{\pi}{2}$) as,

$$\mathcal{SA}_{\theta+\frac{\pi}{2}}(x, y) = \sum_{\omega_r} \mathcal{C}_\theta(x_{r_+}, y_{r_+}, \omega_r) \tag{4.1}$$

where,

$$\omega_r = \frac{\pi \times n_{lobes}}{2r} \tag{4.2}$$

The Complex cell response, $\mathcal{C}_\theta(x_{r_+}, y_{r_+}, \omega_r)$, is computed as explained in Section 5.3.1.3, as the square root of the sum of squared responses of Simple Even and Odd cell responses. But, here it is computed with a different set of parameters, σ_{SA}, γ_{SA}, ω_r instead of σ, γ and ω that we used in Section 5.3.1.3 respectively. The values of σ_{SA}, γ_{SA}, ω_r and other relevant parameters are listed in Table 4.1. Filter size is equal to $2r$ and r is the perpendicular distance between the point, (x, y) at which SA is being computed and the centers of Even and Odd Simple cells. The centers of Even and Odd cells, hence the Complex cell are all located at (x_{r_+}, y_{r_+}), from where the complex cell responses are pooled to compute SA. The term, n_{lobes} in Eq 4.2 determines the number of lobes in the Even and Odd Simple cells. It could be 2 or 4 for even symmetric Simple cells and 3 or 5 for odd symmetric Simple cells. The location from which Complex cell responses are pooled, (x_{r_+}, y_{r_+}) is computed as,

$$x_{r_+} = x + r\cos(\theta + \frac{\pi}{2}) \tag{4.3}$$

$$y_{r_+} = y + r\sin(\theta + \frac{\pi}{2}) \tag{4.4}$$

Similarly, SA at the same location, (x, y), but for the opposite side of border at the same orientation, θ is computed as,

$$\mathcal{SA}_{\theta-\frac{\pi}{2}}(x, y) = \sum_{\omega_r} \mathcal{C}_\theta(x_{r_-}, y_{r_-}, \omega_r) \tag{4.5}$$

where,

$$x_{r_-} = x + r\cos(\theta - \frac{\pi}{2}) \tag{4.6}$$

$$y_{r_-} = y + r\sin(\theta - \frac{\pi}{2}) \tag{4.7}$$

So, for every location on the OB there will be two SA cells capturing the intensity and texture gradients on the two sides abutting the border. It has to be noted that the major axis orientation of the Gabor filters (*i.e.* Even and Odd cells) is same as the local edge orientation, θ. This is because we want to capture the variation of spectral power in a direction orthogonal to the OB, which is captured by the Complex cells with their orientation parallel to the OB.

Let us denote the side of the OB located in the $\theta + \frac{\pi}{2}$ direction as s_+ and the other side, located in the $\theta - \frac{\pi}{2}$ direction as s_-. $\mathcal{SA}_{\theta+\frac{\pi}{2}}(x, y)$ and $\mathcal{SA}_{\theta-\frac{\pi}{2}}(x, y)$ are the SA values

Parameter	**Value**
Min Filter Size	9
Max Filter Size	25
Filter Size Increment Step	2
Aspect Ratio (γ_{SA})	0.8
n_{lobes} (Simple Even cells, S_e)	4
n_{lobes} (Simple Odd cells, S_o)	5
Std dev (Gaussian) (σ_{SA})	0.6r

Table 4.1: Parameters values related to the Simple (Eqs 5.9 and 5.10) and Complex (Eq 5.11) cells used in Spectral Anisotropy computation

corresponding to sides, s_+ and s_-, respectively.

Figure-ground classification at each location is done, as in Chapter 2, Eq 2.6, using a simple classification rule,

$$s_+ := \begin{cases} \text{figure} & \text{if} \quad \mathcal{SA}_{\theta+\frac{\pi}{2}}(x,y) > \mathcal{SA}_{\theta-\frac{\pi}{2}}(x,y) \\ \text{ground} & \text{if} \quad \mathcal{SA}_{\theta-\frac{\pi}{2}}(x,y) \geq \mathcal{SA}_{\theta+\frac{\pi}{2}}(x,y) \end{cases} \tag{4.8}$$

4.3 Data and Methods

The figure-ground dataset, a subset of BSDS 300 dataset, consists of 200 images of size 321×481 pixels, where each image has two ground truth figure-ground labels [106] and corresponding boundary maps. For each image, the two sets of figure ground labels are annotated by users other than those who outlined the boundary maps. The figure-ground ground truth boundary consists of figure side of the boundary marked by *+1* and the ground side boundary by *-1*.

The figure-ground classification accuracy (FGCA) for an image we report is the percentage of the total number of boundary pixels in the ground truth figure/ground label

map for which a correct figure/ground classification decision is made by the SA computation method. Even though SA can be computed at every location where $\mathcal{C}_\theta(x, y, \omega_r)$ is non-zero, the SA responses are compared only at those locations for which ground truth figure/ground labels exist. We report the average of FGCAs of all 200 images here. Whenever the two ground truth label maps differ for the same image, average of the FGCA for both ground truth label maps is reported. Since different figure-ground labelers interpret figure and ground sides differently depending on the context, such differences arise, as a result the self-consistency between figure-ground labelings between the two sets of ground truth annotations is 88%, which is the maximum achievable FGCA for the dataset. At each pixel, the direction of figure, as determined by the model can be correct or wrong. So, the average FGCA for the entire dataset, at chance is 50%, assuming figure/ground relations at neighboring pixels are independent. This assumption is consistent with previously reported results [106], where same assumption was made. The complete details of the figure-ground dataset can be found in [11, 106, 132].

4.4 Results

Color images were converted to Gray scale as before [115], which are used in all the computations below. Gabor filters of size, $9 \times 9, 11 \times 11, \ldots, 25 \times 25$ pixels modeling the Simple Even and Odd cells are used. Their responses are combined as explained in Equation 5.11 to get the contrast invariant Complex cell responses. The number of lobes[1] for Even cells was 5 and Odd cells was 4. The number of orientations was chosen to be 8, equally spaced

[1]Note that Simple Even cell has odd number of lobes and *vice-versa*. "Even" and "Odd" refer to the symmetry of Gabor filters, see Section 1.1.1.

Parameter	Value
Min Filter Size (pixels)	9×9
Max Filter Size (pixels)	25×25
Filter Size Increment (pixels)	2
Number of lobes in Simple Even cells	5
Number of lobes in Simple Odd cells	4
Aspect ratio of simple cells (γ_{SA})	0.8
Cell response type used	Complex
Number of orientations	8
Classification Accuracy (Average)	**62.49**%

Table 4.2: The parameters used in the computation of biologically plausible SA and the FGCA.

between 0 to π. The parameter γ_{SA}, which controls the skewness of the filters was chosen to be 0.8. As explained in Section 4.2, SA is computed for each side of the OB by summing the Complex cell responses of various scales. Figure/ground classification decision at every boundary pixel in the ground truth figure-ground map is made as detailed in Eq 4.8. FGCA for an image is computed as explained in Section 4.3.

The overall FGCA for all 200 images was 62.49%. This FGCA is comparable to the FGCA we obtained with Discrete Fourier Transform (DFT) based method of Chapter 2 (see Section 2.5.3), where it was 62.57% for 1475 randomly chosen boundary locations from 300 BSDS images. This the main result of this chapter. The parameter values used in this case are summarized in Table 4.2. In [106], Ren *et al.* get 55.6% correct figure-ground classification accuracy on the same dataset of 200 images by using size/convexity as a local cue. Even though local contour convexity is a well established local cue [133], it is a weaker cue compared to SA in natural images. This is confirmed based on both DFT based (Chapter 2) and biologically plausible SA computation results as the FGCA is above 62% in both cases.

In the next few paragraphs, we study the robustness of the method to different parameters such as the number of orientations, the skewness parameter γ_{SA} of the filters, number of lobes and also the variation in FGCA when the figure and ground sides are shifted from the border by $1, \ldots, 3$ pixels.

Even though some cells in V1 have 5 or 6 lobes depending on whether the cell is Even or Odd symmetric, a majority of the respective cells have only 3 or 2 lobes. Cells with 5 or 6 lobes capture higher spatial frequencies compared to those with 2 or 3 lobes for the same filter size. In Experiment 2, we intend to verify whether SA can be computed effectively with cells having 2 or 3 lobes, which are more frequently found in V1. So we use Simple Even and Odd cells with 3 and 2 lobes, respectively. Complex cell response is computed as before. Filters of 11 different spatial scales ranging from $7 \times 7, 9 \times 9, \ldots, 25 \times 25$ pixels are used. The number of orientations is 8 and $\gamma_{SA} = 0.8$, same as before. With these parameters, the average FGCA for all 200 images is 60.74%. We see a small drop from 62.49% to 60.74% in the FGCA, but the method performs very well confirming that Simple Even and Odd cells typically found in V1, with 2 or 3 lobes, can also be used to compute SA.

Next we change the number of orientations to 6, equally spaced between 0 and π. Keeping all other parameters same as in the second experiment, we get an average FGCA of 60.68% for the 200 images. When the number of orientations is reduced to 4 with all other parameters being same as in Experiment 2, the average FGCA remains essentially same at 60.51%. These results indicate that the method is quite robust and only 4 orientations are sufficient to get satisfactory FGCA.

In experiment 5, we change the skewness parameter γ_{SA}, which controls the spatial aspect ratio of the filters to 1.0 from 0.8. This makes the filters circular, instead of elongated. With all other parameters being the same as in previous experiment (*i.e.* , 4 orientations, 2 and 3 lobes in Odd and Even symmetric Simple cells respectively and filter sizes, $7 \times 7, 9 \times 9, \ldots, 25 \times 25$), we see very little change in FGCA, which is now 60.33%. So, the skewness parameter has little effect on FGCA. In summary, the FGCA is high and changes very little for different number of orientations, filter sizes and spatial aspect ratios.

To see what the most parsimonious model of the biologically plausible SA computation can capture and to what extent it can explain the results obtained with the multi-scale pooling of Complex cell responses, we stripped down the model to a bare minimum. We use Simple (2 and 3 lobes for Odd and Even cells respectively) and Complex cells of a single scale, with filter size 9×9 pixels, number of orientations was 4, spatial aspect ratio parameter, $\gamma_{SA} = 0.8$ and we obtain the average FGCA of 59.32% for 200 images. The highest FGCA obtained with multi-scale pooling of Complex cell responses from 9 scales was $\approx 62.49\%$. Even the most parsimonious version of SA computation gets a FGCA, which is $\approx 5\%$ less compared to that of the full model (Table 4.2). Also, the SA computation with only a single scale still obtains a higher FGCA than the size/convexity based local cue in [106], which is 55.6%. Given that a single scale Complex cell responses can explain a large proportion of our results is particularly surprising, which indicates the high predictive power of biological SA in determining figure/ground relations accurately. Moreover, in the first experiment where we obtained an FGCA of 62.49%, we used Simple Even and Odd cells having 5 and 4 lobes respectively. So those cells were selective to higher spatial frequency

than in the case of the most parsimonious model.

Lastly, we study how FGCA changes when SA is computed from regions that are slightly shifted away from the OB on both sides. This study is similar to what we did with SVM model in the previous chapter. But in Chapter 3, the dataset consisted of only a few hundred patches selected from random locations on the OB from BSDS dataset, as only one-third of 1475 patches were used. We wanted to see if the results summarized in Table 3.2 still hold, when all the boundary locations in the same database are used. Hence we do the same analysis here, but for all boundary locations of the figure-ground ground truth images.

As in Section 3.5, we were concerned about possible artifactual results due to slight but systematic misalignment between the actual figure-ground border and the human-generated contour which is used to label the ground truth figure/ground relations. So, we shifted the RF centers of Simple and Complex cells by $1, \ldots, 3$ pixels. This is done by modifying Eq 4.3 as, $x^{'}_{r_+} = x + (r + ps)\cos(\theta + \frac{\pi}{2})$ and Eq 4.4 as, $y^{'}_{r_+} = y + (r + ps)\cos(\theta + \frac{\pi}{2})$, where ps denotes the number of pixels shifted away from the OB, which is between $1, \ldots, 3$ pixels. Similarly, Eqs 4.6 and 4.7 are modified to get $x^{'}_{r_-}$ and $y^{'}_{r_-}$ respectively. Complex cell responses are pooled from $(x^{'}_{r_+}, y^{'}_{r_+})$, instead of (x_{r_+}, y_{r_+}) to get $\mathcal{SA}_{\theta+\frac{\pi}{2}}(x, y)$. Similar computation is done for for the other BO direction. In this case, the Gabor filter parameter values listed in Table 4.2 were used, with $ps = 1, 2, 3$ pixels in each case where FGCA of all 200 images is computed. Computing SA for different amounts of pixel shifts did not change the FGCA very much for the same set of 200 images, see Table 4.3. This indicates, as before, in Section 3.5, SA is a property of the figure surface and not an artifact due to misaligned human drawn contours.

Pixels Shifted	FGCA
0	62.49%
1	62.47%
2	62.38%
3	62.22%

Table 4.3: FGCA with Complex cell RFs shifted by 1, 2 and 3 pixels (see text).

4.5 Discussion

We show that pooling of Complex cell responses of multiple scales, whose major axes are parallel to the OB, from a small neighborhood near OB is sufficient to explain how SA may be computed biologically by V1 neurons. Since, on figure side, due to surface curvature, there is spatial compression of uniform features, pooling of only one orientation is sufficient to capture the higher orthogonal spectral power, hence the effect of SA. The model achieves 62.49% FGCA for the 200 images in the database, which is similar to the DFT based SA computation results of Chapter 2, where image patches selected randomly along the OB were used.

A large number of cells (≈50%) in V2 have properties similar to V1 cells [44], motivating the use of Gabor filters to model both V1 and V2 cells. As a result, this particular computation can be carried out in either area V1 or V2 of the visual cortex.

The concept of frequency pooling for a single orientation but different scales is nothing new. Willmore et al. [44], proposed this as a mechanism to detect illusory contours in V2, the same mechanism also serves the purpose of SA computation when the RF is centered on a region close to OB, which is novel.

When a neuron's response is determined jointly by the presence of features of some ori-

entation and the absence of features of certain other orientations, it is said to exhibit tuned suppression. SA, which is the abundance of spectral power of one orientation (orthogonal to figure-ground boundary) compared to parallel orientation, can also be explained on the basis of tuned suppression [44, 134] found in V2 neurons.

Another striking feature of the this biologically plausible SA computation is that even the most parsimonious model with Even and Odd cells of a single scale can explain a large portion of the gain in FGCA, as we obtain 59.32% FGCA with single scale SA computation compared to 62.49% FGCA of the most complex model (Table 4.2). Moreover the model is robust to variations in filter size, aspect ratio and the number of orientations, as detailed in Section 4.4.

Spectral Anisotropy, computed in a biologically plausible manner, as explained here is incorporated into a model of FGO with both local and global cues, which is the focus of next chapter.

4.6 Conclusion

We demonstrate SA can be computed in a biologically plausible manner by pooling Complex cell responses at various scales from small regions on both sides of the OB. The single scale SA computation achieves 59.32% FGCA, whereas the multi-scale computation with 9 scales achieves 62.49% FGCA for all boundary locations of images in the BSDS figure-ground ground truth dataset. This FGCA is nearly the same as the FGCA we saw in [89], where a DFT based method was used for SA computation on a database of image patches, selected from random locations on the OB. We also show the results are not sensitive to

variation in the number of orientations, scales, aspect ratio of the filters or the number of lobes in the filters.

Chapter 5

A figure-ground organization model with local and global cues

5.1 Introduction

A variety of local and global cues mediate the process of FGO, which have been identified by many researchers [3]. The neural mechanism of FGO called Border Ownership coding, role played by Gestalt cues in FGO, their classification into local and global cues, some examples of each type, have been covered in Chapter 1. In this chapter, we present a neurally motivated, feed-forward computational model of FGO incorporating both local and global cues. While we do not attempt to exactly mimic the neural processing mechanism at every step, we try to keep it as biologically motivated as possible.

The contour fragments forming an object's boundary are detected by Simple and Complex cells in the area V1 of primate visual cortex with their highly localized, retinotopically

organized receptive fields (Section 1.1.1). Cells in area V2, which receive input from V1 Complex cells, code for BO by preferentially firing at a higher rate when the figural object is located on the preferred side of the BO coding neuron at its preferred orientation, irrespective of local contrast [40]. Recently, Williford and von der Heydt [81], remarkably show for the first time, that V2 neurons maintain the same BO preference properties even for objects in complex natural scenes, which provides the biological motivation to build a neurally inspired FGO model.

Many computational models [135, 136, 137] have been proposed to explain the neural mechanism by which FGO or BO coding is achieved in the visual cortex. Based on the connection mechanism, those models can be classified as feed-forward [25], feedback [26] or lateral interaction models [137]. The FGO model we develop is a feed-forward model, based on the model of Russell et al. [25], which was developed to explain visual saliency [138] through proto-objects. Our model has three independent feature channels, Color, Intensity and Orientation, similar to that of Russell et al. [25]. The main computational construct of the model is a BO computation mechanism that embodies Gestalt principles of convexity, surroundedness and parallelism, which is identical to all feature channels. Even though the BO computation part is similar to that of Russell et al. [25], we introduce many new modifications to make it suitable for performing FGO and to incorporate local cues, as detailed in Section 5.3. The model, applicable to any natural image, is tested on the widely used BSDS figure/ground dataset. First, we show that even the model with only global cues, devoid of any local cues achieves impressive results on the BSDS figure/ground dataset. Let us call this the *reference model*, against which we compare the performance of models with

added local cues.

We add two local cues to the reference model, Spectral Anisotropy [89] and T-Junctions. The motivation behind adding local cues is their relatively low computational cost compared to global cues. Spectral Anisotropy (SA) was shown to be a valid cue for FGO in Chapters 2 and 4, achieving $> 60\%$ accuracy in predicting which side of an occlusion boundary is figure and which the background. Moreover, SA can be computed efficiently in a biologically plausible way with fast, Fourier domain convolutions (See Chapter 4 for biologically plausible computation of SA), making it an attractive candidate. T-Junctions are commonly viewed as one of the strongest, most unambiguous cues of occlusion and their computation can be explained on the basis of end-stopped cells [12, 139, 140]. This is the biological motivation to incorporate T-Junctions into the model. But the T-Junction computation methods that we employ (Section 5.4.2) do not have a known neural analogue.

At present, only two local cues have been incorporated into our model. Both influence the Orientation channel only as the properties they capture are more closely related to this feature. Certainly, many more local and global cues would be needed to fully solve the Figure/Ground segregation problem in real world images. But, here we proceed with the motivation to investigate how these local cues can be successfully incorporated into the reference model. Second, our purpose is to verify whether local cues can co-exist along with the global cues. If so, how useful are these local cues? Can they lead to a statistically significant improvement in the model's performance when added alone? Finally, are these local cues mutually facilitatory leading to even further improvement, when added together? For these purposes, the minimalistic model with few global cues and even fewer local cues

added to only one of the three feature channels provides an excellent analysis framework.

In Section 5.3, we show how the proto-object based saliency model of Russell et al. [25] is adapted for the purpose of FGO and how new local cues are added. The excellent results of the reference model are first discussed in Section 5.6, then in Sections 5.6.1 and 5.6.2, we show adding local cues individually can lead to statistically significant improvement in the model's performance. We conclude after showing in Section 5.6.4 that co-existence of both local cues leads to even higher improvement in the model's performance.

5.2 Related Work

FGO has been an active area of research in Psychology since nearly a century [7]. Excellent reviews about the Gestalt principles of FGO and grouping can be found in [3, 4]. It is an active area of research in neuroscience [40, 141, 142, 143] and computer vision [106, 144, 145] as well. We limit our literature review to computational models only. Even though the terms "FGO", "BO" or "grouping" are not used in many publications we reviewed, the common goal in all of them is related to inferring depth ordering of objects.

Grossberg and Mingolla [146], Grossberg [147] propose that a reciprocal interaction between a Boundary Contour System (BCS) extracting edges and a Feature Contour System (FCS) extracting surfaces achieves not only FGO, but also 3D perception. In one of the early attempts [148], a two layer network with connections between "computational units" within and across layers is proposed. These units integrate bottom-up edge input with top-down attention input to realize FGO. A model of contour grouping and FGO was proposed in [139] central to which is a "grouping" mechanism. The model not only

generates figure-ground labels, but also simulates the perception of illusory contours. Another influential model was proposed in [149] with feedback and feedforward connections having 8 different computational modules to obtain representations of contours, surfaces and depth. Roelfsema et al. [136], Jehee et al. [150] propose multilayer feedback networks resembling the neural connection pattern in the visual cortex to perform BO assignment through feedback from higher areas. Li Zhaoping *et al.* [151, 152] propose a model of FGO based on V1 mechanisms. The model consists of orientation selective V1 neurons which influence surrounding neurons through mono-synaptic excitatory and di-synaptic suppressive connections. The excitatory lateral connections mimic co-linear excitation [153] and cross-orientation facilitation [154], while inhibitory connections model the iso-orientation suppression [155]. In a related model [137], neurons in V2 having properties of convexity preference, good continuation and proximity was presented. A BO coding model which detects curvatures, L-Junctions and sends proportional signals to a BO layer was proposed by Kikuchi and Akashi [156], where BO signals are propagated along the contour for two sides of BO. The model proposed by Craft et al. [135] consists of edge selective cells, BO cells and multi-scale grouping (G) cells. The G cells send excitatory feedback to those BO cells that are co-circular and point to the center of the annular G cell receptive field. The model incorporates Gestalt principles of convexity, proximity and closure and has T-junctions as local cues. Several models [25, 26, 157, 158, 159] with similar computational mechanisms have been proposed to explain various phenomena related to FGO, saliency, spatial attention, *etc.* A model akin to [135] was proposed in [160], where in addition to G cells the model consists of region cells at multiple scales. In a feedback model [161] based

on the interaction between dorsal and ventral streams, surfaces which are of smaller size, greater contrast, convex, closed, having higher spatial frequency are preferentially determined as figures. The model also accounts for figure-ground cues such as lower region and top-bottom polarity. A 3-layer feed-forward spiking neural model [162] having 2 feature specific channels with excitatory connections between retinotopically organized neurons of different layers was proposed. In a series of papers [127, 163, 164] Sakai and colleagues formulate a BO model in which localized, asymmetric surround modulation is used to detect contrast imbalance, which then leads to FGO.

Differentiation/Integration for Surface Completion (DISC) model [165] was proposed in which BO is computed by detecting local occlusion cues such as T- and L- junctions and comparing non-junction border locations with junction locations for BO consistency with the local cues. A Bayesian belief network based model was proposed [166] in which local cues (curvature and T-junctions) interact with medial axis or skeleton of the shape to determine BO.

A local shapeme based model employing Conditional Random Fields (CRF) to enforce global consistency at T-junctions was proposed in [106]. Hoiem *et al.* [144, 167] used a variety of local region, boundary, Gestalt and depth based cues in a CRF model to enforce consistency between boundary and surface labels achieving impressive results on different datasets. An optimization framework [168] to obtain a 2.1D sketch by constraining the "hat" of the T-junction to be figure and "stem" to be ground was proposed, which uses human labeled contours and T-junctions. In an extension [169], a reformulated optimization over regions, instead of pixels, was proposed. By using various cues such as, curve and junc-

tion potentials, convexity, lower-region, fold/cut and parallelism, Leichter and Lindenbaum [170] train a CRF model to enforce global consistency and produce good results on BSDS figure/ground dataset. In a series of papers Palou and Salembier [171, 172, 173] show how image segmentation and depth ordering (FGO) can be performed using only low-level cues. Their model uses Binary Partition Trees (BPT) [174] for hierarchically representing regions of an image, performs depth ordering by iteratively pruning the branches of BPT enforcing constraints based on T-junctions and other depth related cues. In a recent work [145], which uses Structured Random Forests (SRF) for boundary detection, simultaneous boundary detection and figure-ground labeling is performed. They use shape cues, extremal edge basis functions [15], closure, image torque [175] *etc* to train the SRFs.

Yu et al. [176] present a hierarchical Markov Random Field (MRF) model incorporating rules for continuity of depth on surfaces, discontinuity at edges between surfaces and local cues such as T- and L-junctions. The model learns from a couple examples and effectively does depth segregation, thereby FGO. In [177], a neurally plausible model integrating multiple figure-ground cues using belief propagation in Bayesian networks with leaky integrate and fire neurons was proposed. A simultaneous segmentation and figure-ground labeling algorithm was reported in [178] which uses Angular Embedding [179] to influence segmentation cues from figure-ground cues and *vice-versa*. Similar attempts with primary goal of segmenting images and labeling object classes using figure-ground cues can be seen in [180, 181].

5.3 Model Description

The model consists of three independent features channels, Color, Intensity and Orientation. The features are computed at multiple scales to achieve scale invariance. Orientation selective V1 Simple and Complex cells [182] are excited by edge fragments of objects within their receptive field (Figure 5.1). Let us denote the contrast invariant response of a Complex cell at location (x, y) by $\mathcal{C}_\theta(x, y, \omega)$, where θ is the preferred orientation of the cell and ω is the spatial frequency. The spatial frequency ($\omega = 1.57$, see Table 5.1 for all parameters of the model) is same of all edge responsive cells in our model, except when explicitly stated otherwise. Hence, we omit this variable for the most part, except in Section 5.4.1. Each active Complex cell, $\mathcal{C}_\theta(x, y)$ activates a pair of BO cells, one with a BO preference direction, $\theta + \frac{\pi}{2}$ (a 90° counter-clockwise rotation with respect to θ) denoted as $\mathcal{B}_{\theta+\frac{\pi}{2}}(x, y)$, and the other with $\theta - \frac{\pi}{2}$ BO preference, denoted as $\mathcal{B}_{\theta-\frac{\pi}{2}}(x, y)$. When we talk about the BO response related to a specific figure/ground cue, be it local or global, a subscript is added to the right of the variable. For example, $\mathcal{B}_{\theta-\frac{\pi}{2},TJ}(x, y)$ would be used to denote the BO response related to T-Junctions. Likewise, when specifying scale is necessary, it is denoted by superscript, k. For example, $\mathcal{C}^k_\theta(x, y)$ denotes Complex cell response for orientation θ at location, (x, y) and scale, k. On the other hand, when we need to explicitly specify the feature we talk about, a subscript is added to the left of the variable. For example, ${}_C\mathcal{B}_{\theta-\frac{\pi}{2}}(x, y)$ represents the BO response for the Color feature channel. When a specific BO direction, feature, cue, scale or a location is not important, we just refer to them as, $\mathcal{B}$ cells, $\mathcal{C}$ cells, *etc.* Same applies in all such situations.

Without the influence of any local or global cues, the responses of both BO cells at a

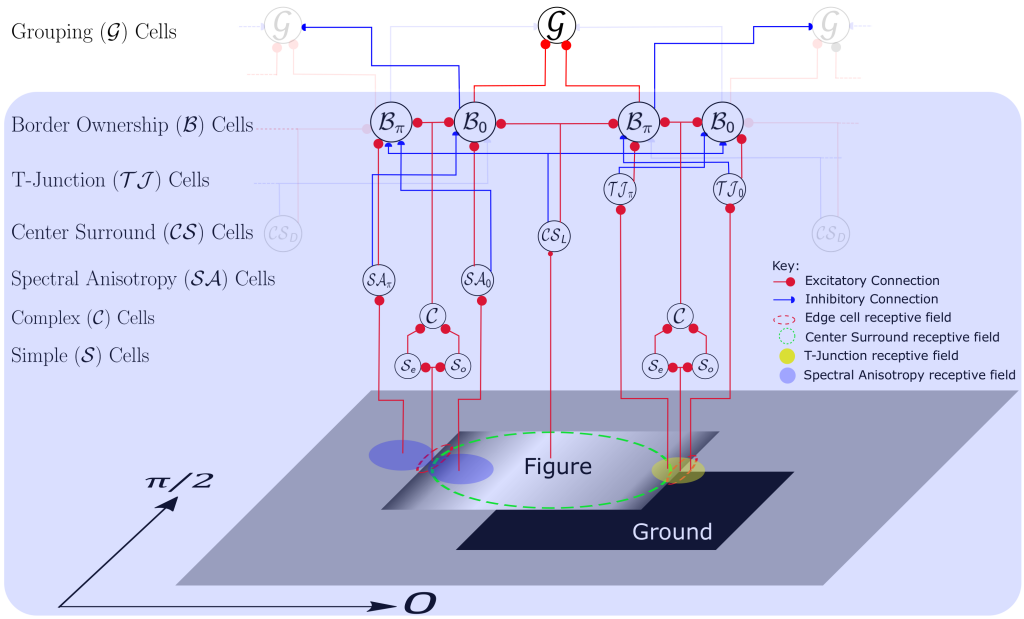

Figure 5.1: Figure-Ground Organization model with local cues: Input to the model are two overlapping squares. Bright foreground square has intensity gradient along the border (vertical orientation), which partially overlaps the black square forming T-Junctions. Network architecture for a single scale and single orientation, $\theta = \frac{\pi}{2}$ shown, but it is same for all 10 scales and 8 orientations. Spectral Anisotropy ($\mathcal{SA}_{\theta\pm\frac{\pi}{2}}$) and T-Junctions ($\mathcal{TJ}_{\theta\pm\frac{\pi}{2}}$) are the new local cues added to the model of [25]. $\mathcal{SA}$ and $\mathcal{TJ}$ cells are active only for the Orientation feature channel, as these are properties related only to that feature. Both $\mathcal{SA}$ and $\mathcal{TJ}$ cells excite $\mathcal{B}$ cells on the same side of the border and inhibit on the opposite side. $\mathcal{TJ}$ cue is computed such that $\mathcal{TJ}$ cells pointing to "stem" of T-Junction are zero, but have a high value for the opposite BO direction. The grouping ($\mathcal{G}$) cells are shown only to emphasize that the model is able to construct proto-objects as in [25] for saliency computation. In this work, our focus is only on the computations upto $\mathcal{B}$ cells (light blue background) for the purpose of FGO. Adapted from Figure 3 of Russell et al. [25]

location will be equal, hence the figure direction at that location is arbitrary. The center-surround cells, denoted as $\mathcal{CS}$ cells, bring about global scene context integration by modulating the $\mathcal{B}$ cell activity. The $\mathcal{CS}_L$ cells (Figure 5.1) extract light objects on dark background, while $\mathcal{CS}_D$ cells code for dark objects on light background. Without the influence of local cues, this architecture embodies the Gestalt properties of convexity, surroundedness and parallelism.

The local cues (see Section 5.4 for computational details of local cues) modulate $\mathcal{B}$ cell activity additionally. Similar to $\mathcal{B}$ cells, a pair of Spectral Anisotropy cells exist for the two opposite BO preference directions at each location, which capture local texture and shading gradients (see Section 5.4.1 for SA computation) on the two sides of the border. Let us denote by $\mathcal{SA}_{\theta+\frac{\pi}{2}}(x,y)$ the cell capturing Spectral Anisotropy for $\theta+\frac{\pi}{2}$ BO direction, likewise $\mathcal{SA}_{\theta-\frac{\pi}{2}}(x,y)$ for the opposite BO direction. The T-Junction cells (see Section 5.4.2 for computational details) also come in pairs, for the two opposite BO directions. Similar to $\mathcal{SA}$ cells, $\mathcal{TJ}_{\theta\pm\frac{\pi}{2}}(x,y)$ hold the T-Junction cue information for the two antagonistic BO directions, $\theta \pm \frac{\pi}{2}$. Both these type of cells excite $\mathcal{B}$ cells of the same BO direction and inhibit the opposite BO direction $\mathcal{B}$ cells. For example, $\mathcal{SA}_{\theta+\frac{\pi}{2}}(x,y)$ excites $\mathcal{B}_{\theta+\frac{\pi}{2}}(x,y)$ and inhibits $\mathcal{B}_{\theta-\frac{\pi}{2}}(x,y)$.

The influence of $\mathcal{CS}$ cells, $\mathcal{SA}$ cells and $\mathcal{TJ}$ cells on $\mathcal{B}$ cells is controlled by a set of weights (not shown in Figure 5.1). Local cues are active in the Orientation channel only. The interplay of all these cues leads to the emergence of figure/ground relations strongly biased for one of the two BO directions at each location. The $\mathcal{B}$ cell activity, modulated by both local and global cues, can be integrated by the grouping cells ($\mathcal{G}$ cells, Figure 5.1)

to compute proto-objects and saliency (See Russell et al. [25] for details). But, our main focus in this work is on FGO, hence we restrict our computation only upto $\mathcal{B}$ cells (blue shaded rectangle in Figure 5.1). The network architecture depicted in Figure 5.1 is the same computational construct that is applied at every scale, for every feature and orientation. The successive stages of model computation are explained in the following subsections.

5.3.1 Computation of feature channels

We consider Color, Intensity and Orientation as three independent feature channels in our model, the computation of each is described in the following sections.

5.3.1.1 Intensity channel

The input image consists of Red (r), Blue (b) and Green (g) color channels. The intensity channel, I is computed as average of the three channels, $I = (r+b+g)/3$. As with all other feature channels, a multi-resolution image pyramid is constructed from the intensity channel (Section 5.3.2). The multi-resolution analysis allows incorporation of scale invariance into the model.

5.3.1.2 Color opponency channels

The color channels are first normalized by dividing each r, g or b value by I. From the normalized r, g, b channels, four color channels, Red ($\mathcal{R}$), Green ($\mathcal{G}$), Blue ($\mathcal{B}$) and Yellow ($\mathcal{Y}$) are computed as,

$$\mathcal{R} = \max\left(0, r - \frac{g+b}{2}\right) \tag{5.1}$$

$$\mathcal{G} = \max\left(0, g - \frac{r+b}{2}\right) \tag{5.2}$$

$$\mathcal{B} = \max\left(0, b - \frac{g+r}{2}\right) \tag{5.3}$$

$$\mathcal{Y} = \max\left(0, \frac{r+g}{2} - \frac{|(r-g)|}{2} - b\right) \tag{5.4}$$

In Eq 5.4, the symbol, | | denotes absolute value.

The four opponent color channels, $\mathcal{RG}$, $\mathcal{GR}$, $\mathcal{BY}$ and $\mathcal{YB}$ are computed as,

$$\mathcal{RG} = \max(0, \mathcal{R} - \mathcal{G}) \tag{5.5}$$

$$\mathcal{GR} = \max(0, \mathcal{G} - \mathcal{R}) \tag{5.6}$$

$$\mathcal{BY} = \max(0, \mathcal{B} - \mathcal{Y}) \tag{5.7}$$

$$\mathcal{YB} = \max(0, \mathcal{Y} - \mathcal{B}) \tag{5.8}$$

5.3.1.3 Orientation channel

The Orientation channel is computed using the canonical model of visual cortex [182], where quadrature phase, orientation selective, Gabor kernels are used to model the V1 simple cells. The responses of Simple cells are non-linearly combined to obtain the contrast invariant, orientation selective response of the Complex cell. Mathematically, the receptive fields of even and odd symmetric Simple cells can be modeled as the cosine and sine components of a complex Gabor function - a sinusoidal carrier multiplied by a Gaussian envelope. The RF of a Simple Even cell, $s_{e,\theta}(x, y)$ is given by,

$$s_{e,\theta}(x, y) = e^{-\frac{X^2+\gamma^2 Y^2}{2\sigma^2}} \cos(\omega X) \tag{5.9}$$

where, $X = x\cos(\theta) + y\sin(\theta)$ and $Y = -x\sin(\theta) + y\cos(\theta)$ are the rotated coordinates, σ is the standard deviation of the Gaussian envelope, γ is the spatial aspect ratio (controlling how elongated or circular the filter profile is), ω is the spatial frequency of the cell and θ is the preferred orientation of the simple cell. Similarly, the receptive field of a Simple Odd cell is defined as,

$$s_{o,\theta}(x, y) = e^{-\frac{X^2+\gamma^2 Y^2}{2\sigma^2}} \sin(\omega X) \tag{5.10}$$

Simple even and odd cells responses, respectively denoted $S_{e,\theta}(x, y)$ and $S_{o,\theta}(x, y)$ are computed by correlating the intensity image, $I(x, y)$ with the respective RF profiles. The Complex cell response, $\mathcal{C}_\theta(x, y)$ is calculated as,

$$\mathcal{C}_\theta(x,y) = \sqrt{S_{e,\theta}(x,y)^2 + S_{o,\theta}(x,y)^2} \tag{5.11}$$

Eight orientations in the range, $[0,\pi]$, at intervals of $\frac{\pi}{8}$ are used.

5.3.2 Multiscale pyramid decomposition

Let us denote a feature map, be it Orientation ($\mathcal{C}_\theta$), Color ($\mathcal{RG}$, $\mathcal{BY}$, *etc*) or Intensity feature map, at image resolution by a common variable, $\beta^0(x,y)$. The next scale feature map, $\beta^1(x,y)$ is computed by downsampling $\beta^0(x,y)$. The downsampling factor can be either $\sqrt{2}$ (half-octave) or 2 (full octave). Bi-linear interpolation is used to compute values in the down-sampled feature map, $\beta^1(x,y)$, which is the same interpolation scheme used in all cases of up/down sampling. Similarly, any feature map $\beta^k(x,y)$ of a lower scale, k is computed by downsampling the higher scale feature map, $\beta^{k-1}(x,y)$ by the appropriate downsampling factor. As the numerical value of k increases, the resolution of the map at that level in the pyramid decreases. The feature pyramids thus obtained are used to compute BO pyramids explained the next section.

In addition to the multiscale pyramids of independent feature channels, we compute the multiscale local cue pyramids for SA and T-Junctions as well. To denote the local cue map at a specific scale, as with feature pyramids, the scale parameter k is used. For example, $\mathcal{SA}^k_{\theta+\frac{\pi}{2}}(x,y)$ denotes the Spectral Anisotropy feature map for $\theta+\frac{\pi}{2}$ border ownership direction at scale, k. Similarly T-Junction pyramids at different scales for $\theta \pm \frac{\pi}{2}$ BO directions are denoted by $\mathcal{TJ}^k_{\theta\pm\frac{\pi}{2}}(x,y)$. The local cue pyramids are computed by successively downsampling the local cue maps at native resolution, $\mathcal{SA}_{\theta\pm\frac{\pi}{2}}$ and $\mathcal{TJ}_{\theta\pm\frac{\pi}{2}}$ (see Section 5.4 for

their computation details).

Whereas the multi-resolution feature pyramids for Color, Intensity and Orientation are computed by downsampling the image to a lower resolution first and recomputing the features at each scale, the local cue pyramids are computed differently. The local cues of SA and T-Junctions are computed *only once* at the native image resolution, but the local cue maps are repeatedly downsampled such that they have similar data structure as the feature pyramids. With similar data structure, the local cues can be more easily integrated into the model with minimal modification in the overall model architecture. When we do a multi-resolution feature pyramid decomposition, it allows us to keep the filter kernel size of $\mathcal{B}$ cells constant and small. This reduces the number of mathematical operations compared to having $\mathcal{B}$ cells of different radii and keeping the image size constant. To ascertain the computational cost (See Sections 5.6.3 and B.3 for computational cost analysis) of adding local cues to the model is minimal, we enforce local cues to have the same data structure as feature pyramids. This enables us to take advantage of the fixed size convolutional kernels for $\mathcal{B}$ cells.

5.3.3 Border Ownership pyramid computation

The operations performed on any of the features ($\mathcal{C}_\theta$ or I) or the sub-type of features like $\mathcal{RG}, \mathcal{BY}$ is the same. BO responses are computed by modulating $C_\theta(x, y)$ by the activity of center-surround feature differences on either sides of the border, as in Russell et al. [25]. Each feature map, $\beta^k(x, y)$, is correlated with the center-surround filters to get center-surround ($\mathcal{CS}$) difference feature pyramids. Two types of center-surround filters, $cs_{on}(x, y)$

(ON-center) and $cs_{off}(x, y)$ are defined as,

$$cs_{on}(x,y) = \frac{1}{2\pi\sigma_{in}^2} e^{-\frac{(x^2+y^2)}{2\sigma_{in}^2}} - \frac{1}{2\pi\sigma_{out}^2} e^{-\frac{(x^2+y^2)}{2\sigma_{out}^2}} \tag{5.12}$$

$$cs_{off}(x,y) = -\frac{1}{2\pi\sigma_{in}^2} e^{-\frac{(x^2+y^2)}{2\sigma_{in}^2}} + \frac{1}{2\pi\sigma_{out}^2} e^{-\frac{(x^2+y^2)}{2\sigma_{out}^2}} \tag{5.13}$$

where $\sigma_{out}, \sigma_{in}$ are the standard deviations of the outer and inner Gaussian kernels respectively.

The center-surround dark pyramid, $\mathcal{CS}_D^k$ is obtained by correlating the feature maps, β^k with the $cs_{off}(x, y)$ filter followed by half-wave rectification,

$$\mathcal{CS}_D^k(x,y) = \max(0, \beta^k(x,y) * cs_{off}(x,y)) \tag{5.14}$$

which detects weak/dark features surrounded by strong/light ones. In Eq 5.14, the symbol, $*$ denotes 2D correlation [183]. Similarly, to detect strong features surrounded by weak background, a $\mathcal{CS}_L^k$ pyramid is computed as,

$$\mathcal{CS}_L^k(x,y) = \max(0, \beta^k(x,y) * cs_{on}(x,y)) \tag{5.15}$$

The $\mathcal{CS}$ pyramid computation is performed this way for all feature channels except for the Orientation channel. For the Orientation feature channel, feature contrasts are not typically symmetric as in the case of other features, but oriented at a specific angle. Hence, the $cs_{on}(x, y)$ and $cs_{off}(x, y)$ filter kernels in Equations 5.14 and 5.15 are replaced by even

symmetric Gabor filters, $s_{e,\theta}(x,y)$ (ON-center) and $-s_{e,\theta}(x,y)$ (OFF-center) of opposite polarity respectively. But, in this case, different set of parameter values are used. Instead of $\gamma = 0.5$, $\sigma = 2.24$ and $\omega = 1.57$ used in Section 5.3.1.3, here we use $\gamma_1 = 0.8$, $\sigma_1 = 3.2$ and $\omega_1 = 0.7854$ respectively. The parameter values are modified in this case such that the width of the center lobe of the even Gabor filters (ON and OFF-center) matches the zero crossing diameter of the $cs_{on}(x,y)$ and $cs_{off}(x,y)$ filter kernels in Equations 5.14 and 5.15. As a result, the ON-center Gabor kernel detects bright oriented edges in a dark background, instead of symmetric feature discontinuities detected by $cs_{on}(x,y)$. Similarly, the OFF-center Gabor filter detects activity of dark edges on bright backgrounds.

An important step in BO computation is normalization of the center-surround feature pyramids, $\mathcal{CS}_L^k(x,y)$ and $\mathcal{CS}_D^k(x,y)$. Let $\mathcal{N}_1(\cdot)$ be used to denote the normalization operation, which is same as the normalization used in [138], but done after rescaling $\mathcal{CS}_D$ and $\mathcal{CS}_L$ pyramids to have the same range, $[0,\ldots,M]$. Similarly the local cue pyramids, $\mathcal{SA}_{\theta+\frac{\pi}{2}}^k(x,y)$ and $\mathcal{SA}_{\theta-\frac{\pi}{2}}^k(x,y)$ are also normalized using the same method and in the same range, $[0,\ldots,M]$. In the same way, $\mathcal{TJ}_{\theta\pm\frac{\pi}{2}}^k(x,y)$ pyramids are also normalized. This normalization step enables comparison of different features and local cues on the same scale, hence the combination of feature and local cue pyramids.

Since, we compute BO on the normalized light and dark CS pyramids, $\mathcal{N}_1(\mathcal{CS}_L^k(x,y))$ and $\mathcal{N}_1(\mathcal{CS}_D^k(x,y))$ separately and combine them at a later stage, let us denote, the corresponding BO pyramids by $B_{\theta\pm\frac{\pi}{2},L}^k(x,y)$ and $B_{\theta\pm\frac{\pi}{2},D}^k(x,y)$ respectively. We explain the BO pyramid computation for $B_{\theta+\frac{\pi}{2},L}^k(x,y)$ and $B_{\theta+\frac{\pi}{2},D}^k(x,y)$ which have a BO preference direction of $\theta+\frac{\pi}{2}$. Computation of $B_{\theta-\frac{\pi}{2},L}^k(x,y)$ and $B_{\theta-\frac{\pi}{2},D}^k(x,y)$ is analogous.

Let $\hat{K}_{\theta+\frac{\pi}{2}}(x,y)$ denote the kernel responsible for mapping the object activity from normalized $\mathcal{CS}_L$ and $\mathcal{CS}_D$ pyramids to the object edges, which is implemented with von Mises distribution. von Mises distribution is a normal distribution on a circle [184]. The unnormalized von Mises distribution, $K_{\theta+\frac{\pi}{2}}(x,y)$ is defined as [25],

$$K_{\theta+\frac{\pi}{2}}(x,y) = \frac{\exp\left[(\sqrt{x^2+y^2}-R_0)\sin(\tan^{-1}(\frac{y}{x})-(\theta+\frac{\pi}{2})\right]}{I_0(\sqrt{x^2+y^2}-R_0)} \tag{5.16}$$

where $R_0 = 2$ pixels is the radius of the circle on which the von Mises distribution is defined, $\theta+\frac{\pi}{2}$ is the angle at which the normal distribution is concentrated [184] on the circle (also called mean direction), and I_0 is the modified Bessel function of the first kind. The distribution is then normalized as,

$$\hat{K}_{\theta+\frac{\pi}{2}}(x,y) = \frac{K_{\theta+\frac{\pi}{2}}(x,y)}{\max\left(K_{\theta+\frac{\pi}{2}}(x,y)\right)} \tag{5.17}$$

$\hat{K}_{\theta-\frac{\pi}{2}}(x,y)$ is computed analogously.

The BO pyramid, $B^k_{\theta+\frac{\pi}{2},L}(x,y)$ for light objects on dark background capturing the BO activity for $\theta+\frac{\pi}{2}$ direction is computed as,

$$\begin{aligned}\mathcal{B}^k_{\theta+\frac{\pi}{2},L}(x,y) = \max\Bigg(0, \mathcal{C}^k_\theta(x,y) \times \Bigg(1 + \bigoplus_{j\geq k}\frac{1}{2^j}\hat{K}_{\theta+\frac{\pi}{2}}(x,y) * \mathcal{N}_1(\mathcal{CS}^j_L(x,y)) \\ - w_{opp}\bigoplus_{j\geq k}\frac{1}{2^j}\hat{K}_{\theta-\frac{\pi}{2}}(x,y) * \mathcal{N}_1(\mathcal{CS}^j_D(x,y))\Bigg)\Bigg)\end{aligned} \tag{5.18}$$

Similarly, the BO pyramid for $\theta+\frac{\pi}{2}$ direction for a dark object on light background is obtained by correlating normalized $\mathcal{CS}$ maps with $\hat{K}_{\theta\pm\frac{\pi}{2}}$ and summing the responses for all

scales greater than the scale, k at which BO map is being computed as,

$$\mathcal{B}^k_{\theta+\frac{\pi}{2},D}(x,y) = \max\Bigg(0, \mathcal{C}^k_\theta(x,y) \times \Bigg(1 + \bigoplus_{j\geq k} \frac{1}{2^j}\hat{K}_{\theta+\frac{\pi}{2}}(x,y) * \mathcal{N}_1(\mathcal{CS}^j_D(x,y)) \\ -w_{opp} \bigoplus_{j\geq k} \frac{1}{2^j}\hat{K}_{\theta-\frac{\pi}{2}}(x,y) * \mathcal{N}_1(\mathcal{CS}^j_L(x,y))\Bigg)\Bigg) \tag{5.19}$$

where, w_{opp} is the synaptic weight for the inhibitory signal from the $\mathcal{CS}$ feature map of opposite contrast polarity. The symbol, $\bigoplus$ is used to denote pixel-wise addition of responses from all scales greater than k, by first up-sampling the response to the scale at which $\mathcal{B}^k_{\theta+\frac{\pi}{2},D}(x,y)$ is being computed. The other two pyramids, $\mathcal{B}^k_{\theta-\frac{\pi}{2},L}(x,y)$ and $\mathcal{B}^k_{\theta-\frac{\pi}{2},D}(x,y)$ for the opposite BO direction are computed analogously.

With the BO pyramids related to dark and light $\mathcal{CS}$ pyramids already computed, we turn our attention to the computation of the local cue related BO pyramids. The local cue pyramids at different scales, $\mathcal{SA}^k_{\theta\pm\frac{\pi}{2}}(x,y)$ and $\mathcal{TJ}^k_{\theta\pm\frac{\pi}{2}}(x,y)$ are constructed, as explained in Sections 5.4.1 and 5.4.2, by successively down-sampling the local cue maps computed at native image resolution. Both local cues excite $\mathcal{B}$ cells of the same BO direction and inhibit the opposite BO direction $\mathcal{B}$ cells.

The BO pyramid for $\theta + \frac{\pi}{2}$ BO direction related to the local cue, SA denoted as, $\mathcal{B}^k_{\theta+\frac{\pi}{2},SA}(x,y)$ is computed as,

$$\mathcal{B}^k_{\theta+\frac{\pi}{2},SA}(x,y) = \max\Bigg(0, \mathcal{C}^k_\theta(x,y) \times \Bigg(1 + \bigoplus_{j\geq k} \frac{1}{2^j}\hat{K}_{\theta+\frac{\pi}{2}}(x,y) * \mathcal{N}_1(\mathcal{SA}^j_{\theta+\frac{\pi}{2}}(x,y)) \\ -w_{opp} \bigoplus_{j\geq k} \frac{1}{2^j}\hat{K}_{\theta-\frac{\pi}{2}}(x,y) * \mathcal{N}_1(\mathcal{SA}^j_{\theta-\frac{\pi}{2}}(x,y))\Bigg)\Bigg) \tag{5.20}$$

where we can see the SA cell ($\mathcal{SA}^k_{\theta+\frac{\pi}{2}}(x,y)$) having same BO preference as the BO cell, $\mathcal{B}^k_{\theta+\frac{\pi}{2},SA}(x,y)$ has an excitatory effect on the BO cell, but $\mathcal{SA}^k_{\theta-\frac{\pi}{2}}(x,y)$ has an inhibitory effect. The synaptic weight, w_{opp} remains unchanged as in Eqs 5.18 and 5.19. The BO pyramid, $\mathcal{B}^k_{\theta-\frac{\pi}{2},SA}(x,y)$ related to SA, for opposite BO direction is computed in the same way.

The BO pyramid related to T-Junctions for the BO direction, $\theta+\frac{\pi}{2}$ is computed as,

$$\begin{aligned}\mathcal{B}^k_{\theta+\frac{\pi}{2},TJ}(x,y) = \\ \max\Bigg(0, \mathcal{C}^k_\theta(x,y) \times \Bigg(1 + \bigoplus_{j\geq k}\frac{1}{2^j}\hat{K}_{\theta+\frac{\pi}{2}}(x,y) * \mathcal{N}_1(\mathcal{TJ}^j_{\theta+\frac{\pi}{2}}(x,y)) \\ - w_{opp}\bigoplus_{j\geq k}\frac{1}{2^j}\hat{K}_{\theta-\frac{\pi}{2}}(x,y) * \mathcal{N}_1(\mathcal{TJ}^j_{\theta-\frac{\pi}{2}}(x,y))\Bigg)\Bigg)\end{aligned} \tag{5.21}$$

The corresponding T-Junction pyramid for the opposite BO direction, $\theta-\frac{\pi}{2}$, denoted as $\mathcal{B}^k_{\theta-\frac{\pi}{2},TJ}(x,y)$ is computed analogously.

The combined BO pyramid for direction, $\theta+\frac{\pi}{2}$ is computed by summing global and local cue specific BO pyramids as,

$$\begin{aligned}\mathcal{B}^k_{\theta+\frac{\pi}{2}}(x,y) = \alpha_{ref}\Big(\mathcal{B}^k_{\theta+\frac{\pi}{2},L}(x,y) + \mathcal{B}^k_{\theta+\frac{\pi}{2},D}(x,y)\Big) \\ + \alpha_{SA}\Big(\mathcal{B}^k_{\theta+\frac{\pi}{2},SA}(x,y)\Big) + \alpha_{TJ}\Big(\mathcal{B}^k_{\theta+\frac{\pi}{2},TJ}(x,y)\Big)\end{aligned} \tag{5.22}$$

where α_{ref}, α_{SA} and α_{TJ} are weights such that $\alpha_{ref}+\alpha_{SA}+\alpha_{TJ}=1$, that control the contribution of $\mathcal{CS}$, $\mathcal{SA}$ and $\mathcal{TJ}$ cues to the BO response at that location respectively. By setting the weights to 0 or 1, we can study the effect of individual cue on BO response. It should be noted that the local cues are active only for the Orientation channel, so for

the other channels, α_{SA} and α_{TJ} will be set to zero, by default. In the absence of local cues, combination of light and dark BO pyramids (first term in Eq 5.22) results in contrast polarity invariant BO response. The corresponding BO pyramid for opposite BO preference, $\mathcal{B}^k_{\theta-\frac{\pi}{2}}(x,y)$ is computed as in Eq 5.22 by summing the light, dark and local cue BO pyramids of opposite BO preference.

Since the BO responses, $\mathcal{B}^k_{\theta\pm\frac{\pi}{2}}(x,y)$, are computed for each orientation, θ there will be multiple BO responses active at a given pixel location. But the boundary between figure and ground can only belong to the figure side, *i.e.* there can only be one winning BO response for a given location. So, the winning BO response, denoted as $\widehat{\mathcal{B}}^k_{\theta+\frac{\pi}{2}}(x,y)$ is computed as,

$$\widehat{\mathcal{B}}^k_{\theta+\frac{\pi}{2}}(x,y) = \begin{cases} \max\left(0, \mathcal{B}^k_{\theta+\frac{\pi}{2}}(x,y) - \mathcal{B}^k_{\theta-\frac{\pi}{2}}(x,y)\right), & \text{if } \theta = \widehat{\theta} \\ 0, & \text{otherwise} \end{cases} \tag{5.23}$$

where $\widehat{\theta} = arg\max_\theta(\left|\mathcal{B}^k_{\theta+\frac{\pi}{2}}(x,y) - \mathcal{B}^k_{\theta-\frac{\pi}{2}}(x,y)\right|)$ is the orientation for which absolute difference between antagonistic pair of BO responses is maximum over all orientations. This gives the edge orientation at that location. So, the winning BO pyramid, $\widehat{\mathcal{B}}^k_{\theta+\frac{\pi}{2}}$ has non-zero response at a location only if the difference between the corresponding pair of BO responses for $\widehat{\theta}$ is non-negative. The winning BO pyramid, $\widehat{\mathcal{B}}^k_{\theta-\frac{\pi}{2}}$ for the opposite direction is computed analogously.

Upto this point, the computation for all feature channels is identical. Now, if we denote the feature specific winning BO pyramid for $\theta+\frac{\pi}{2}$ direction for the Color channel by ${}_C\widehat{\mathcal{B}}^k_{\theta+\frac{\pi}{2}}$, Intensity feature channel by ${}_I\widehat{\mathcal{B}}^k_{\theta+\frac{\pi}{2}}$ and Orientation feature channel by ${}_O\widehat{\mathcal{B}}^k_{\theta+\frac{\pi}{2}}$, then the final BO map, $\widetilde{\mathcal{B}}_{\theta+\frac{\pi}{2}}(x,y)$ for $\theta+\frac{\pi}{2}$ BO direction is computed by linearly combining the

up-sampled feature specific BO maps across scales as,

$$\widetilde{\mathcal{B}}_{\theta+\frac{\pi}{2}}(x,y) = \bigoplus_{k=1}^{N_s} \left({}_C\widehat{\mathcal{B}}^k_{\theta+\frac{\pi}{2}}(x,y) + {}_I\widehat{\mathcal{B}}^k_{\theta+\frac{\pi}{2}}(x,y) + {}_O\widehat{\mathcal{B}}^k_{\theta+\frac{\pi}{2}}(x,y) \right) \tag{5.24}$$

where $\bigoplus$ represents pixel-wise addition of feature specific BO responses across scales after up-sampling each map to native resolution of the image. Similarly, $\widetilde{\mathcal{B}}_{\theta-\frac{\pi}{2}}$ is computed for $\theta - \frac{\pi}{2}$ BO direction. As we can see in Eq 5.24, the contribution of every feature channel to the final BO map is the same, *i.e.* , feature combination is equally weighted. Ten spatial scales ($N_s = 10$) are used. All parameters of the model are summarized in Table 5.1. In the end, we get 16 BO maps at image resolution, 8 each for $\theta + \frac{\pi}{2}$ and $\theta - \frac{\pi}{2}$ BO directions respectively.

5.4 Computation of local cues

5.4.1 Computation of Spectral Anisotropy

Spectral Anisotropy is computed by pooling Complex cell responses of various spatial frequencies from small image regions on either sides of the boundary as explained in Chapter 4. Two SA maps, $\mathcal{SA}_{\theta+\frac{\pi}{2}}(x,y)$ for the BO direction, $\theta + \frac{\pi}{2}$ and $\mathcal{SA}_{\theta-\frac{\pi}{2}}(x,y)$ for the BO direction, $\theta - \frac{\pi}{2}$ are computed for each orientation, θ for the two sides of the border. The SA maps thus obtained are decomposed into multiscale pyramids, $\mathcal{SA}^k_{\theta\pm\frac{\pi}{2}}(x,y)$, where superscript, k denotes scale, by successive downsampling, which are used to compute the cue specific BO pyramids as explained in Section 5.3.3, Equation 5.20.

5.4.2 Determining T-Junctions

The object edges and the regions bound by those edges called "segments" are obtained using the gPb+ucm+OWT image segmentation algorithm [185], referred to as the gPb algorithm in other parts of this thesis. Image segmentation, partitioning of an image into disjoint regions, is considered as a pre-processing step occurring prior to FGO. The edges obtained using the gPb algorithm are normalized to have values in $[0, 1]$ range. Edge strength, in $[0, 1]$ range, is proportional to the confidence with which the algorithm finds an actual object edge in the image. Any edges having a value smaller than the threshold of 0.41 are removed. This threshold was tuned using the training set of images. From the thresholded set of edges a Contour Map as shown in Figure 5.2 (B) and a Segmentation Map as shown in Figures 5.2 (C), respectively are obtained. The Contour Map has the same edge information, except that the pieces of contours appearing to meet at a junction location are uniquely numbered. The Segmentation Map contains uniquely numbered disjoint regions bound by the contours. The Contour Map and Segmentation Maps are just a convenient way of representing the same edge information. Only the locations at which exactly 3 distinct contours meet in the Contour Map (Figure 5.2 (B)) and correspondingly the locations at which exactly 3 distinct segments meet in the Segmentation Map (Figure 5.2 (C)) are considered for T-Junction determination. Such locations can be easily determined from the Segmentation and Contour maps.

As shown in Figure 5.2 (E) and 3F, at each junction location we have three regions, R_1, R_2 and R_3 and contours, c_1, c_2 and c_3 meeting. At each such junction, a circular mask of N_{mask} pixels is applied and the corresponding small patches of the segmentation map and

contour map are used for further analysis. We determine the contours forming the "hat" of the T-Junction (foreground) and the corresponding figure direction in two different ways: (1) based on the area of regions meeting at junction location within the small circular disk around junction; (2) based on the angle between contours meeting at the junction location. Finally, only those junctions locations for which figure direction, as determined based on both methods, is matching are introduced into the FGO model as T-Junction local cues. Matching based on two different methods improves the overall accuracy in correctly identifying the "hat" (foreground) and "stem" (background) of T-Junctions, in effect the correct figure direction.

The local neighborhood of T-Junction influence is set to be a circular region of radius 15 pixels. All the border pixels near the junction location within a radius of 15 pixels that belong to the "hat" of the T-Junction are set to +1 for the appropriate BO direction. Remember that for each orientation, θ we will have two T-Junction maps, one for the BO preference direction, $\theta+\frac{\pi}{2}$ denoted as $\mathcal{TJ}_{\theta+\frac{\pi}{2}}$ and the other for the opposite BO preference, $\theta-\frac{\pi}{2}$ denoted as $\mathcal{TJ}_{\theta-\frac{\pi}{2}}$. A pixel in $\mathcal{TJ}_{\theta+\frac{\pi}{2}}(x,y)$ is set to +1 if the direction of figure, as determined by both methods (Sections 5.4.2.1 and 5.4.2.1) is $\theta+\frac{\pi}{2}$, *i.e.* "stem" of the T-Junction is in the $\theta-\frac{\pi}{2}$ direction. Similarly, $\mathcal{TJ}_{\theta-\frac{\pi}{2}}(x,y)$ computed. The T-Junction maps thus obtained are decomposed into multiscale pyramids, $\mathcal{TJ}^{k}_{\theta\pm\frac{\pi}{2}}(x,y)$, where superscript, k denotes scale, by successive downsampling, which are used to compute the cue specific BO pyramids as explained in Section 5.3.3, Equation 5.21.

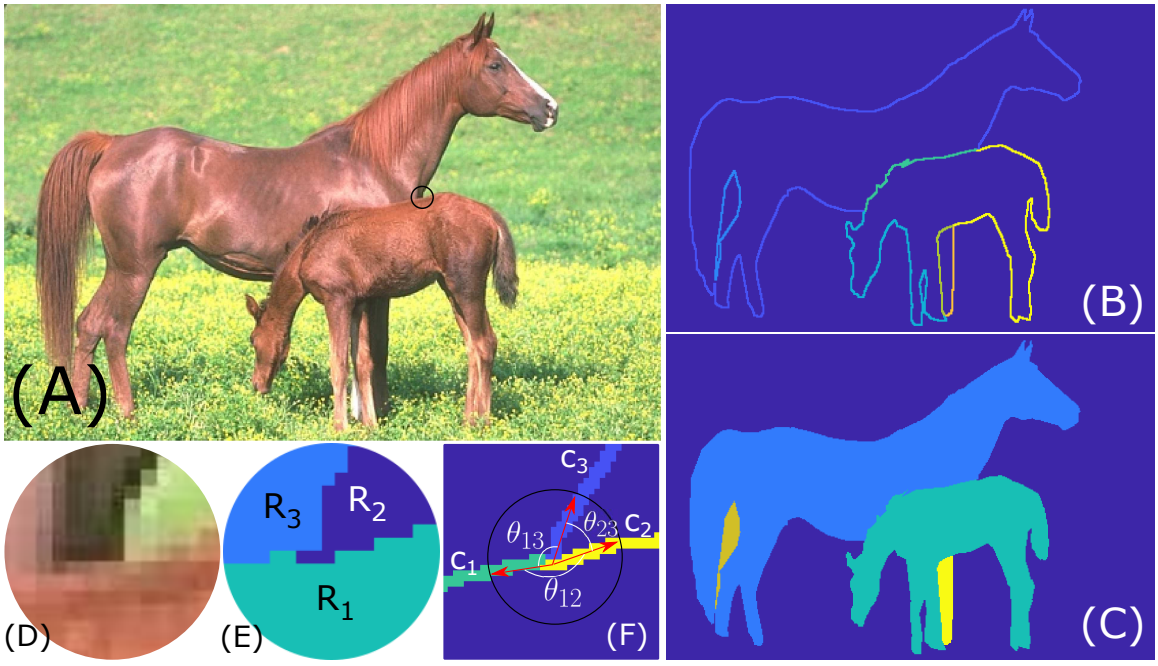

Figure 5.2: T-Junctions: Image (A) with T-Junction (black circle), the corresponding contours (B) and segments (C) are shown. Area based T-Junction determination: In (D), a small patch from image used for determining T-Junctions is shown. (E) Areas of three regions, R_1, R_2 and R_3 meeting at the T-Junction are determined. Contours abutting the segment (R_1) with largest area form the "hat" of T-Junction. Angle based T-Junction determination: (F) From the junction location, 7 pixels are tracked for each contour, c_1, c_2 and c_3. Three vectors (red arrows) are defined based on the start (always junction location) and end points for each contour. The angles between the three vectors are determined. Contours for which largest angle (θ_{12}) is observed form the "hat" of the T-Junction. Only matching T-Junctions based on segment area and contour angle are used in the model

5.4.2.1 Area based T-Junction determination

Let R_1, R_2 and R_3 be the three regions at a junction location (x, y) (Figure 5.2E). After extracting the circular region around the junction by applying a circular mask of radius, 6 pixels, we count the number of pixels belonging to each of the regions, R_i. The region, R_i having the largest pixel count is determined as the figure region. In Figure 5.2E, R_1 is the region with largest pixel count, hence determined as the foreground. The contours abutting the figure region, R_1 as determined by pixel count, which are c_1 and c_2 (Figure 5.2F), form the "hat" of the T-Junction. Contour c_3 forms the "stem" of the T-Junction, which belongs to the background.

The local orientation at each contour location is known. Vectors of length $1 - 3$ pixels, normal to the local orientation are drawn at each "hat" contour location within the 15×15 pixel neighborhood. If the normal vector intersects the figure region, R_1, as determined based on region area, the edge/contour location is given a value of $+1$ in the T-Junction map for the appropriate BO direction, which can be $\theta + \frac{\pi}{2}$ or $\theta - \frac{\pi}{2}$. This is done for every pixel in the edge/contour map within a neighborhood of 15 pixel radius around the T-Junction location for those contours (c_1 and c_2) that form the "hat" of the T-Junction. For example, in Figures 5.2E and 3F, if the local orientation of c_1 and c_2 is roughly 0, then the end point of normal vector in the $\theta - \frac{\pi}{2}$ direction intersects with the figure region, R_1, as determined based on the segment area. So, the T-Junction map for $\theta - \frac{\pi}{2}$ BO preference direction is set to $+1$ within the circular neighborhood of 15×15 pixels. The T-Junction map for $\theta + \frac{\pi}{2}$ BO direction will be zero.

5.4.2.2 Angle based T-Junction determination

In this method, as in Section 5.4.2.1, a small circular patch of radius, 7 pixels is extracted from the contour map around T-junction location. Pixels belonging to each contour, c_i meeting at the junction are labeled with a distinct number, so for each contour, c_i we track the first 7 pixels starting from the junction location. Since the starting point for each contour, c_i is the same, the total angle at junction location is 360^0. For each contour, we define a vector (red arrows in Figure 5.2 F) from the junction location to the last tracked point on the contour. We then compute the angle between the vectors corresponding to contours. The contours between which angle is the largest form the "hat" of the T-junction. For example, in Figure 5.2F, θ_{12} is the angle between c_1 and c_2, which is also the largest of the three angles, θ_{12}, θ_{32} and θ_{13}. So, in the angle based T-junction computation also, c_1 and c_2 are determined to form the "hat" of the T-junction. The figure direction at every pixel of the "hat" contours is determined as in Section 5.4.2.1.

Among all the potential T-Junctions determined using the angle based method, potential Y-Junctions and Arrow junctions are discarded based on the angle formed by the contours at junction location. If the largest angle is greater than 180^0, such junctions are discarded. Since the largest angle greater than 180^0 is typically seen in the case of Arrow-junctions, we do not include them in the computation. Arrow junctions appear in a scene when the corner of a 3D structure is seen from outside. In the same way, if each angle at a junction location is within $120^0 \pm 10^0$, such junctions are discarded as those are most likely Y-Junctions. Y-Junctions appear in a scene when a 3D corner is viewed from inside the object, for example, corner of a room viewed from inside the room. Rest of the T-Junctions are included in our

computation. Angle based filtering of potential Arrow or Y-Junctions was not considered in previous methods [171, 173].

T-Junctions and their figure directions are determined using both Segment Area based and Contour Angle based methods and the T-Junctions are incorporated into the model only in those cases, where both methods give matching figure direction. This reduces the overall T-Junctions in the model, but since each method is prone to some inaccuracies, using matched T-junctions reduces the number of false positives. Since, not only detecting the T-Junction locations, but determining the figure side (*i.e.* the "hat" of the T-junction) is also very critical in FGO, having high agreeableness between both methods was found to be very important.

Accurately determining the figure side of a T-junction from a small neighborhood of 10-15 pixels is quite challenging because within that small neighborhood we generally do not have any information to give away figure/ground relations, other than contour angle and segment area. Even though key point detection is a well studied area, hence locating a T-Junction is not problematic, deciding which of the three region is the foreground based on information from a 7×7 pixel neighborhood is extremely challenging. So, when locally determining figure side of a T-junction, segment area and contour angle were found to be the most exploitable properties.

5.5 Data and methods

We use the same BSDS figure/ground ground truth dataset for evaluation of our model that we used in Chapter 4, Section 4.3, with a few differences.

The figure-ground classification accuracy (FGCA) for an image we report is the percentage of the total number of boundary pixels in the ground truth figure/ground label map for which a correct figure/ground classification decision is made by the model described in Section 5.3. Even though the model computes BO response at every location where $\mathcal{C}_\theta$ cells are active, the BO responses are compared only at those locations for which ground truth figure/ground labels exist.

As in Chapter 4, whenever the two ground truth label maps differ for the same image, average of the FGCA for both ground truth label maps is reported. To remind the readers, the self-consistency between figure-ground labelings between the two sets of ground truth annotations is 88%, which is the maximum achievable FGCA for the dataset. Also, the chance level FGCA of the dataset, as in Chapter 4, is 50%, assuming figure/ground relations at neighboring pixels are independent.

Different from Chapter 4, where the average FGCA for all 200 images were reported, here the BSDS figure/ground dataset is randomly split into training set of 100 images and test set of 100 images. Parameters of the model are tuned for the training dataset and the optimal values of parameters found for the training set are used to evaluate the FGCA of the test set of images. The average FGCA that we report for the entire test set is the average of FGCAs of all 100 images in the test set. Results are reported for test set of images only.

5.6 Results and Discussion

To remind the readers, the model with only global cues of convexity, surroundedness and parallelism, without any local cues is referred to as the reference model. As explained in Section 5.3, local cues, SA and T-Junctions are added to the Orientation feature channel of the reference model. As we have previously described in Section 5.3.3, by setting $\alpha_{SA} = 0$ and $\alpha_{TJ} = 0$ in Eq 5.22, the model with local cues can be reduced to the reference model. Similarly, by switching the weights for each local cue to zero, the effect of the other local cue on FGO can be studied. As explained in Section 5.3.3, the winning BO pyramids are up-sampled to image resolution and summed across scales and feature channels (Eq 5.24) for each BO direction to get the response strength for that BO direction. The BO information derived this way is compared against the ground-truth from BSDS figure/ground dataset.

First, we wanted to quantify the performance of the reference model, which is devoid of both local cues, in terms of FGCA. With $\alpha_{SA} = 0$ and $\alpha_{TJ} = 0$, the overall FGCA for 100 test images was 58.44% (standard deviation = 0.1146). With only global cues, the 58.44% FGCA we achieved is 16.88% above chance level (50%). Hence, we can conclude that the global Gestalt properties of convexity, surroundedness and parallelism, which the reference model embodies, are important properties that are highly useful in FGO. The parameters used in the reference model computation are listed in Table 5.1. Unless stated otherwise explicitly, those parameters in Table 5.1 remain unchanged for the remaining set of results that we are going to discuss. Only the parameters specifically related to the addition of local cues are separately tuned and will be explicitly reported.

Next, we wanted to study the effect of adding each local cue individually (Sections 5.6.1

Parameter	Value
γ	0.5
σ	2.24
ω	1.57
σ_{in}	0.90
σ_{out}	2.70
R_0	2.0
w_{opp}	1.0
σ_1	3.2
γ_1	0.8
ω_1	0.7854
N_s	10

Table 5.1: Parameters of the reference FGO model without any local cues

and 5.6.2) and then the effect of both local cues together (Sections 5.6.4).

5.6.1 Effect of adding Spectral Anisotropy

As explained in Section 5.4.1, Spectral Anisotropy was computed at the native resolution of the image by pooling Complex cell responses at many scales for each orientation. For each orientation, θ, two SA maps, $\mathcal{SA}_{\theta+\frac{\pi}{2}}$ and $\mathcal{SA}_{\theta-\frac{\pi}{2}}$ are created for respective antagonistic BO directions with respect to θ. The SA maps are then decomposed into multiscale pyramids by successively downsampling. The SA pyramids are then incorporated into the model as explained in Eq 5.20 and Eq 5.22. In this case, parameters α_{ref} and α_{SA} are tuned for the training dataset and α_{TJ} is set to 0.

The parameter tuning procedure we use here is the same for other cases as well. We use multiresolution grid search for parameter tuning with the condition that the sum of tuned parameters should be 1. In this case, the condition was $\alpha_{ref} + \alpha_{SA} = 1$. We stop refining the resolution of the grid when the variation in FGCA upto second decimal point is zero,

i.e. , only small changes are seen from third digit onward, after the decimal point.

The optimal parameters were found to be, $\alpha_{ref} = 0.35$ and $\alpha_{SA} = 0.65$ for the training dataset. With these optimal parameter values, the FGCA for the test set was 62.69% (std. dev = 0.1204), which is a 7.3% improvement in the model's performance after adding the local cue, Spectral Anisotropy, compared to the reference model's FGCA of 58.44%. To verify if the improvement in FGCA that we see is statistically significant, we performed an unpaired sample, right tailed t-test (Table 5.2), where the null hypothesis was that the means of FGCAs of the reference model and the model with SA are equal. The alternate hypothesis was that the mean FGCA of the model with SA is higher than that of the reference model. The significance level, $\alpha = 0.05$ was chosen. For other results (Sections 5.6.2, 5.6.4) as well, we do the same type of test, where the reference model's FGCA is compared with that of modified model's FGCA having different local cues. Hereafter, we refer to them as *statistical tests*.

Statistical tests show that the mean FGCA of the model with SA is significantly higher than that of the reference model ($p = 5.2 \times 10^{-301}$). This demonstrates SA is a useful cue and can be successfully incorporated into the reference model, adding which results in statistically significant improvement in the model's performance. This, and all other results are summarized in Table 5.2 for the test dataset.

5.6.2 Effect of adding T-Junctions

As described in Section 5.4.2, T-Junctions are computed at image resolution using the segmentation and edge maps obtained using the gPb [185] algorithm. Each of the T-

Junction maps for the 16 different BO directions is successively downsampled to create multiscale T-Junction pyramids. The T-Junction pyramids are incorporated into the model as explained in Eq 5.21 and Eq 5.22 and by setting $\alpha_{SA} = 0$. The other two parameters, α_{ref} and α_{TJ} are tuned on the training dataset. With optimal parameter values, $\alpha_{ref} = 0.03$ and $\alpha_{TJ} = 0.97$ (and $\alpha_{SA} = 0$), the FGCA for the test set was found to be 59.48% (std. dev. = 0.1127). Compared to the reference model's FGCA of 58.44%, we see that adding T-Junctions derived from gPb [185] based edges, improves the model's performance in terms of FGCA by 1.78%. Based on the statistical tests (Table 5.2), we find that the improvement in FGCA that we see is indeed statistically significant.

Given that T-Junctions are generally thought of as one of the strongest and most unambiguous local cues of occlusion, the relatively small, but statistically significant improvement in FGCA that we see here is not very convincing. Our first thought was that the boundaries that we obtain from the gPb image segmentation algorithm may not be conducive to T-Junction computation. To remove this doubt, we replaced gPb with another highly efficient boundary detection algorithm by Leordeanu et al. [186]. We also tried the CORF push-pull based [187] and Gabor filter bank based edge detection methods. Deriving T-Junctions using the boundaries extracted with these methods did not give any more encouraging results, either. T-Junctions derived from Gabor based edges were least encouraging as the method is sensitive to textures as well. Moreover, accurate orientation estimation at a junction location is a problem common to all the edge detection methods, which has been discussed by Iverson and Zucker [188].

At this point, after trying several algorithms for automatic edge extraction, we were

not able to isolate the exact reason for less than satisfactory improvement in FGCA after adding T-Junctions derived from automatically extracted edges. One reason could be that the quality of boundaries we get from automatic segmentation algorithms may not be good enough for reliable T-Junction estimation. For example, automatically extracted boundaries may not coincide with actual object boundaries in the image, texture markings on object surfaces may be detected as actual boundaries which a human observer would typically ignore, *etc.* Or, the combination of Segment Area and Contour Angle based T-Junction determination method that we are using may be missing many true positive and detecting many false positives. To isolate the problem, we decided to use human labeled boundaries that are provided with the BSDS dataset, instead of automatically extracted boundaries, to derive T-Junctions.

From the human labeled boundaries, we computed T-Junctions, again using the same methods described in Section 5.4.2. So, except for using human labeled boundaries instead of gPb boundaries, all other computation was kept same. Since we are using human labeled boundaries instead of gPb edges, the quality of T-Junctions can be different now. As a result, the reliability of the T-Junction cue introduced into the model can also change. So, we found the optimal values for α_{ref} and α_{TJ} again using multiresolution grid search on the training dataset. With optimal values, $\alpha_{ref} = 0.0125$ and $\alpha_{TJ} = 0.9875$, the new FGCA after adding T-Junctions derived from human labeled boundaries was 61.98% (std. dev = 0.0965) for the test dataset. Compared to the reference model's FGCA of 58.44%, this is a 6.05% improvement with T-Junctions alone, when they are derived from human labeled boundaries instead of boundaries extracted automatically. This is much higher compared to

the 1.78% improvement that we saw when T-Junctions derived from gPb edges were added. As before, statistical tests revealed the improvement in FGCA was statistically significant (Table 5.2).

Even though the 6.05% improvement we observe when T-Junctions derived from human drawn contours is nearly 3 times better compared to the 1.78% FGCA improvement we see when T-Junctions derived from gPb [185] are added, it is still not better than the 7.3% improvement we saw with SA. This is a bit surprising because T-Junctions are generally regarded as a stronger cue of occlusion. But, at the same time, we need to remind ourselves that SA can be computed at every border location of an object, whereas T-Junctions can be computed only at a few sparse location where exactly 3 different regions partially occlude each other. Considering the sparsity of T-junctions, we can say T-Junctions are stronger FGO cues compared to SA, if they can be accurately detected and incorporated. But, almost comparable improvement for both SA and T-Junctions, even T-Junctions they were derived from human drawn contours, made us curious to know if there was some thing else that we have not unraveled so far.

To put these doubts to rest, we decided to use T-Junctions directly taken[1] from the ground truth figure/ground labels. This increased the overall FGCA to 64.77%, a 10.83% improvement compared to that of reference model (details in B.1). This dramatic difference compelled us to take a deeper look at the figure/ground ground truth labels and human marked boundaries, which revealed some interesting anomalies like "inverted" and misla-

[1]It is important to point out that while we computed T-Junctions based on Segment Area and Contour Angle (Sections 5.4.2, 5.4.2.1 and 5.4.2.2) when we derived T-Junctions from human labeled boundaries and gPb algorithm based edges, here T-Junctions are derived directly from figure/ground labels that are provided by the database creators

beled T-Junctions (B.2). Also, in many cases edges that would be typically drawn by human labelers or detected by any edge detection algorithm are edited out of the figure/ground ground truth, more details in B.2. Next, we turn our attention to the computational complexity of adding local cues.

5.6.3 Computational complexity of adding local cues

The most time consuming, computationally intensive part of T-junction computation is the gPb [185] based image segmentation as it involves filtering for brightness, color and texture gradients, multiscale cue combination and oriented watershed transform, which involves large matrix factorization. We utilize this algorithm *as is*, hence we will not delve into exact estimation of computational complexity for this step. Once the contours and segmentation maps are obtained using gPb algorithm, the computation of each T-Junction using both methods described in Section 5.4.2.1 and Section 5.4.2.2 involves multiplying the edge maps, segmentation maps with masks of appropriate sizes, counting and tracking pixels, computing angles, *etc*, which roughly translates into a computational complexity of $O(N_{mask})^2$ for both methods, where $N_{mask} = 13$ pixels for Segment Area based T-Junction computation (Section 5.4.2.1) and $N_{mask} = 15$ pixels for Contour Angle based T-Junction computation (Section 5.4.2.2). Typically $3 - 10$ T-Junctions are found in an image. So, once edges/segmentation map is computed, since only few T-Junctions are typically present in images and the size of mask is not very large, subsequent computation is not very time consuming, even though the complexity is $O(N_{mask})^2$. With appropriate modifications, it should be possible to reduce the computational complexity of T-Junction determination,

which is not optimized at the moment.

The most computationally intensive part of SA computation is the correlations involved in Eq 4.1, but are implemented as convolutions, which has a computational complexity of $O(N_r \times N_c \times \log(N_r \times N_c))$ when implemented in Fourier domain, where N_r and N_c are the number of rows and columns in the image. A more detailed analysis of the computational cost of the entire model with and without local cues is done in Section B.3, where computational cost is quantified in terms of the number of Floating Point Operations (FLOPs) per image. The analysis in Section B.3 shows that with optimized computation, the cost of adding local cues can be moderately low, in the range of 20% - 24%.

Based on the computational complexity of adding T-Junctions, anomalies discussed in B.2, comparing the results in Table 5.2 and from B.1 we can make some important observations. The quality of edges that we get from an automated segmentation algorithm, like gPb [185], hence the T-Junctions derived from them, has a significant impact on the number and correctness of the T-Junctions added to the model. Based on the 6.05% improvement in FGCA we observe in the case of T-Junctions derived from human labeled contours, we can say with better image segmentation algorithms, we should see more benefit from adding T-Junctions. From a computational cost point of view, even though the cost of adding a T-Junction is $O(N_{mask}^2)$, given their sparsity (typically 3-10 T-Junctions per image), adding them as a local cue is justified. We consider image segmentation as a pre-processing step; hence we ignore the computational cost of finding edges. The presence of "inverted" T-Junctions, misinterpreted ground truth and partially missing object boundaries (B.2) is also part of the reason why we do not see as high improvement in FGCA as expected after

adding T-Junctions, in both cases.

Tse and Albert [189] argue that high level surface and volume analysis takes place first, and only after such an analysis, a T-Junction is interpreted to be an occlusion cue. The fact that we see 3 times better FGCA with T-Junctions derived from human labeled boundaries compared to those from gPb based boundaries supports this argument because the increase in FGCA in the former case is a result of fewer false positives. This indicates human observers draw object boundaries after interpreting the global scene structure, hence T-Junctions that are in agreement with global scene structure are retained in the human labeled contours, others even though locally valid, but contradicting with the global context are discarded.

The traditional view that T-Junctions are unambiguous cues of occlusion has also been challenged by psychophysics experiments of McDermott [190], where they find that making occlusion decisions from a small aperture, typically a few pixels wide, in real images is hard for humans too. Several studies also suggest junctions in general, hence T-Junctions, can be cues for image segmentation, but not for occlusion reasoning [191]. The presence of "inverted" T-Junctions, which arise due to properties of objects in the scene, is worth mentioning in this context. Moreover, there is no direct neurophysiological evidence that suggests existence of cells specifically tuned to T-Junctions in the visual cortex [192]. All these previous works and our own results do not support the generally held view that T-Junctions are the most unambiguous occlusion cues. But, these cues are useful, produce statistically significant improvement in FGCA, but not highly unambiguous in all circumstances.

5.6.4 Effect of both Spectral Anisotropy and T-Junctions

In the first case, SA is computed as explained in Section 5.4.1, T-Junctions are computed as explained in Section 5.4.2, where T-Junctions are derived from automatically extracted edges using the gPb algorithm. Both cues are added to the model according to Eq 5.22. The parameters α_{ref}, α_{SA} and α_{TJ} are tuned simultaneously on the training dataset using multiresolution grid search as before, with the constraint, $\alpha_{ref}+\alpha_{SA}+\alpha_{TJ} = 1$. The optimal values of the parameters were found to be, $\alpha_{ref} = 0.05$, $\alpha_{SA} = 0.15$ and $\alpha_{TJ} = 0.80$. All other parameters remained unchanged as shown in Table 5.1. The FGCA of the combined model with both local cues, Spectral Anisotropy and T-Junctions computed from gPb edges was 63.57% (std. dev = 0.1179) for the test dataset, which is higher that the FGCAs we obtained for the individual cues when they were added separately. We see an improvement in FGCA of 8.78% compared to that of the reference model with no local cues. As before, an unpaired sample, right tailed t-test comparing the reference model's figure/ground decisions and the combined model's figure/ground decisions with both SA and T-Junctions, derived from automatically extracted edges showed statistically significant improvement (Table 5.2).

In addition to comparing the performance of the model with both local cues with the Reference model, we also compared the performance of the model with both local cues (Ref model + SA + T-Junctions) to the model with only one (Ref model + SA) local cue. Unpaired sample right-tailed t-tests were used again with a significance level of 0.05. In this case the the null hypothesis is that adding T-Junctions to the Reference Model with SA does not lead to statistically significant improvement in FGCA. The alternate hypothesis is that adding T-Junctions leads to statistically significant improvement in FGCA when

compared to the FGCA of Reference (global cues only) + SA model. Tests show adding T-Junctions to the Reference + SA model leads to a statistically significant improvement ($p = 1.8911 \times 10^{-17}$).

In the next case, we replace T-Junctions derived from automatically extracted edges, with T-Junctions derived from human labeled boundaries. SA is added as before. The optimal values of the parameters for the training dataset were found to be $\alpha_{ref} = 0.0125$, $\alpha_{SA} = 0.0375$ and $\alpha_{TJ} = 0.95$. With these parameter values, the FGCA of the combined model with SA and T-Junctions derived from human labeled contours was 65.48% (std. dev = 0.1170) for the test set of images, a 12.04% improvement over the reference model's FGCA of 58.44%. Once again, statistical tests confirmed that the improvement in FGCA that we see is statistically significant, details in Table 5.2. Again, we compared FGCAs of Reference Model + SA + T-Junctions with that of Reference Model + SA using unpaired sample right-tailed t-tests with a significance level of 0.05. In this case, T-Junctions were derived from human drawn contours. Tests show adding T-Junctions, derived from human drawn contours to the Reference + SA model again leads to a statistically significant improvement ($p = 2.242 \times 10^{-75}$).

In summary, we show that both SA and T-Junctions are useful local cues of FGO, which produce statistically significant improvement in FGCA when added alone. We saw an 8.78% improvement in FGCA of the combined model with both local cues when T-Junctions were added based on the automatically extracted edges and 12.04% improvement when the same were derived from the human drawn contours. This improvement, from only two local cues added to one of the three feature channels is highly impressive. Moreover, the three feature

	FGCA (std. dev)	**%age increase**	**Stat Sig?**	**p-value**
Reference Model	58.44% (0.1146)	-	-	-
With SA	62.69% (0.1204)	7.3%	Yes	5.2×10^{-301}
With T-Junctions (gPb [185] based boundaries)	59.48% (0.1127)	1.78%	Yes	3.38×10^{-26}
With T-Junctions (human labeled boundaries)	61.98% (0.0965)	6.05%	Yes	0
With SA and T-Junctions (gPb [185] based boundaries)	63.57% (0.1179)	8.78%	Yes	0
With SA and T-Junctions (human labeled boundaries)	65.48% (0.1170)	12.04%	Yes	0

Table 5.2: Summary of results for the test dataset: Adding SA to the reference model improves the FGCA by 7.3%. With T-Junctions derived from automatically extracted edges, the FGCA improvement is small compared to when they are derived from human drawn boundaries. Each individual local cue, added alone, produces statistically significant improvement in model performance, in terms of FGCA. When both are added together, the FGCA observed is higher than that we see with individual local cues, indicating the local cues are mutually non-inhibitory. Numbers within parentheses in Column 2 represent the standard deviation in FGCA. All results are statistically significant

channels were weighted equally. The results could have been better, if we had tuned the weights for individual feature channels. But, as we intended to study how to incorporate the local cues and their relative importance in FGO, feature specific weight tuning was not done, but we consider to do this in future. Moreover, it is important to note that FGCA of the model with both local cues is always higher than the FGCAs of models with individual local cues. This suggests the local cues are mutually facilitatory, which is further validated by the fact that we see statistically significant improvement in FGCA when T-Junctions are added as an additional cue to the Reference model having SA as one of the local cues. Figures 5.3 and 5.4 show FGO results for some example images from the test dataset when both SA and T-Junctions added.

Also, as we show in Section B.3, the computational overhead of adding both local cues in an optimized manner is relatively low, in the range of 20% - 24%, yielding $\approx 9\%$ improvement in model's performance. Given that the feature channel weights are unoptimized, local cues added to only one of three feature channels and the model, at present, in not optimized for best FGCA[2], the additional cost of adding local cues to achieve an improvement in FGCA upwards of $\geq 9\%$ is justified.

Our objective in this study was study the benefits of local cues, hence the model parameters were not tuned for best FGCA. So, we compared models with varying number of local cues with the Reference model having only global cues under identical testing conditions. Under identical testing conditions, we surmised we would see the benefits of local cues in comparison to not having them. And we show the local cues are useful. But the FGCA of

[2]See Chapter 8 for a discussion on how FGCA of the model can be improved even with existing local cues

the model can be improved by tuning the inhibitory weight, w_{opp} for each feature and each local cue and tuning feature specific weights in Eq 5.24. In addition, increasing the number of scales in the model, having von Mises kernels, $\mathcal{CS}$ cells and $\mathcal{B}$ cells of multiple radii can all lead to even better FGCA. Having von Mises kernels, $\mathcal{CS}$ cells and $\mathcal{B}$ cells of multiple radii of multiple radii can help capture the convexity and surroundedness cues better. Also, the model's figure-ground response is computed by modulating the activity of $\mathcal{C}_\theta$ cells, which are computed using Gabor filter kernels. The response of $\mathcal{C}_\theta$ cells may not always exactly coincide with human drawn boundaries in the ground-truth, with which we compare the model's response to calculate FGCA. Hence, averaging the BO response in a small 2×2 pixel neighborhood and then comparing that with the ground-truth FG labels could yield improved FGCA. In future, we would like to explore these ideas in order to improve FGCA. Also, we would like to explore global cues such as symmetry, medial axis and color based local cues as FGO is a complex mid-level visual process mediated by several local and global cues.

Our model, not optimized for best performance, still compares well with the existing models of FGO tested on the BSDS figure-ground dataset. Here, we are comparing the FGCA of the model with both local cues, where T-Junctions were automatically extracted from the gPb [185] based edges, with those methods that are fully automated. As we can see in Table 5.3, our current model has better FGCA than the model proposed by Maire [178]. It should be noted that all the methods in Table 5.3 are machine learning based, incorporating a much higher number of cues. While Maire [178] uses 64 different *shapemes*, descriptors of local shape derived from object edges, Ren et al. [106] incorporates empirical

frequencies of 4 different junction types in addition to shapemes in a CRF based model. Also, Ren et al. [106] compare figure/ground relations with the ground-truth only on a partial set of locations where their edge detection algorithm finds a matching edge with the ground-truth. It is not clear what percentage of edges match with the ground-truth. Palou and Salembier [171] use 8 color based cues, in addition to T-Junctions and local contour convexity in their model. The other two ([144, 145]) models use a much larger number of cues to achieve FGO. So, we need to keep in mind we are comparing our model having only two local (SA and T-Junctions) and global (convexity and surroundedness) cues with others having much larger number of cues. Moreover, none of the models we are comparing with is strictly local Gestalt cue based nor neurally motivated. To the best of our knowledge, there are no comparable neurally inspired, fully automated models that are tested on the BSDS figure-ground dataset. Hence, our's is the first such model tested on commonly used FGO database in the Computer Vision community. The model proposed by Sakai et al. [127] is tested on BSDS FG database, but it requires human drawn contours.

We introduce a few novel methods in the computation of the model. First, demonstrating that Spectral Anisotropy can be computed with Simple and Complex cells found in area V1 is a novel contribution. The significance of this computation is that it demonstrates SA can be computed in low level visual areas, even in the striate cortex and it does not require specialized cells to detect these shading/texture gradients. Only a specific arrangement of Complex cells of various spatial frequencies on each side of the border is sufficient. These cues, first mathematically shown to be useful by Huggins et al. [193], were psychophysically validated by Palmer and Ghose [194]. We showed these patterns are abundantly found in

Algorithm	**FGCA**
M. Maire etal, ECCV, 2010 [178]	62%
Our method	**63.5**%
X. Ren etal, ECCV, 2006 [106]	68.9%
P. Salembier etal, IEEE TIP, 2013 [171]	71.3%
CL. Teo, etal, CVPR 2015 [145]	74.7%
D. Hoiem etal, ICCV 2007 [144]	79%

Table 5.3: Comparison of FGCA of our model with existing fully automated FGO models: Our model performs better than Maire [178], which uses 64 different *shapeme* based cues. Ren et al. [106] use empirically measure junction frequencies of 4 different junction types along with *shapeme* cues in a CRF model. They compare with FG ground-truth only on a partial set of edges. Other models use a higher number of cues for FGO. With only a few cues, our model which is Gestalt cue based, neurally motivated and built with the purpose of studying effect of local cues, still performs competitively with existing models. As discussed in Chapter 8, the model's FGCA can be substantially improved with some minimal modifications.

natural images [195] and can be efficiently computed using 1D FFTs [89]. Now, we show that these cues can be computed in a biologically plausible manner, only using Complex cells found commonly in striate cortex. Next, in the computation of T-Junctions, we filter out Y-Junctions and Arrow junctions using the angle property of these junction types. Since Y-Junctions and Arrow junctions are not occlusion cues, ideally those should not be considered as T-Junctions, hence we device a method to remove such junctions. To the best of our knowledge, previous methods [171, 172, 173] that use T-Junctions as FGO cues have not looked closely at this issue, which we consider novel in our approach. Also, we explicitly compute the local figure/ground relations at a T-Junction local based on local information, which is new. And, the way we reorganize the local cue information in such a way that the same computational routine can be used for incorporation of both cues into the model is noteworthy. With this, the implementation is made more efficient, allowing easy parallelization, suitable for Graphics Processing Units (GPU) and other hardware. More-

over, the combination of features and local cues is done at a late stage (Eq 5.22), which allows independent and parallel computation of features and local cues, which again makes the model computationally more efficient and allows parallelization. Plus, the local cues are integrated within the framework of Russell et al. [196] such that no additional modification in the model architecture is necessary for the computation of proto-object based saliency. With the addition of local cues, which result in more reliable BO information, the proto-object based saliency should also improve proportionately.

Lastly, we investigate if the influence of local cues should be strictly local or global. In our model, even though the local cues, SA and T-Junctions, are computed based on the analysis of a strictly local neighborhood around the object boundary, they modulate the activity of $\mathcal{B}$ cells at all scales, *i.e.* , their influence is global in nature. Should the influence of local cues be also local? To answer this question, we added local cues only at the top 2 layers of the model, tuned the optimal parameters, α_{ref}, α_{SA} and α_{TJ} accordingly and recomputed FGCA. We found that with local cue influence at only the top two layers, the FGCA we obtained was lower than having them at all scales (See Appendix B.4 for details). This confirms the influence of local cues need not be local, even though their computation should be strictly local, which is the case in our model. On the other hand, having them influence $\mathcal{B}$ cells of all scales leads to higher FGCA.

5.7 Conclusion

We develop a biologically motivated, feed-forward computational model of FGO with local and global cues. Spectral Anisotropy and T-Junctions are the local cues newly in-

Figure 5.3: Figure/Ground classification results in a few example images: For the images in the first column, the figure/ground ground-truth maps are shown in column 2, where a white pixel denotes the figure side of the border, black pixel, the ground side. Middle column shows the figure/ground classification map for the reference model with no local cues. In images of columns 3–5, if a white pixel on the gray background indicates that a correct figure/ground decision was made by the model at that location, a black pixel indicates it was wrong, in comparison to the ground truth. Column 4 images represent figure/ground classification maps for the model with both local cues, Spectral Anisotropy and T-Junctions, where T-Junctions are derived from automatically extracted edges. Column 5 again shows the figure/ground classification maps for the model with both local cues, but here T-Junctions were derived from human drawn contours

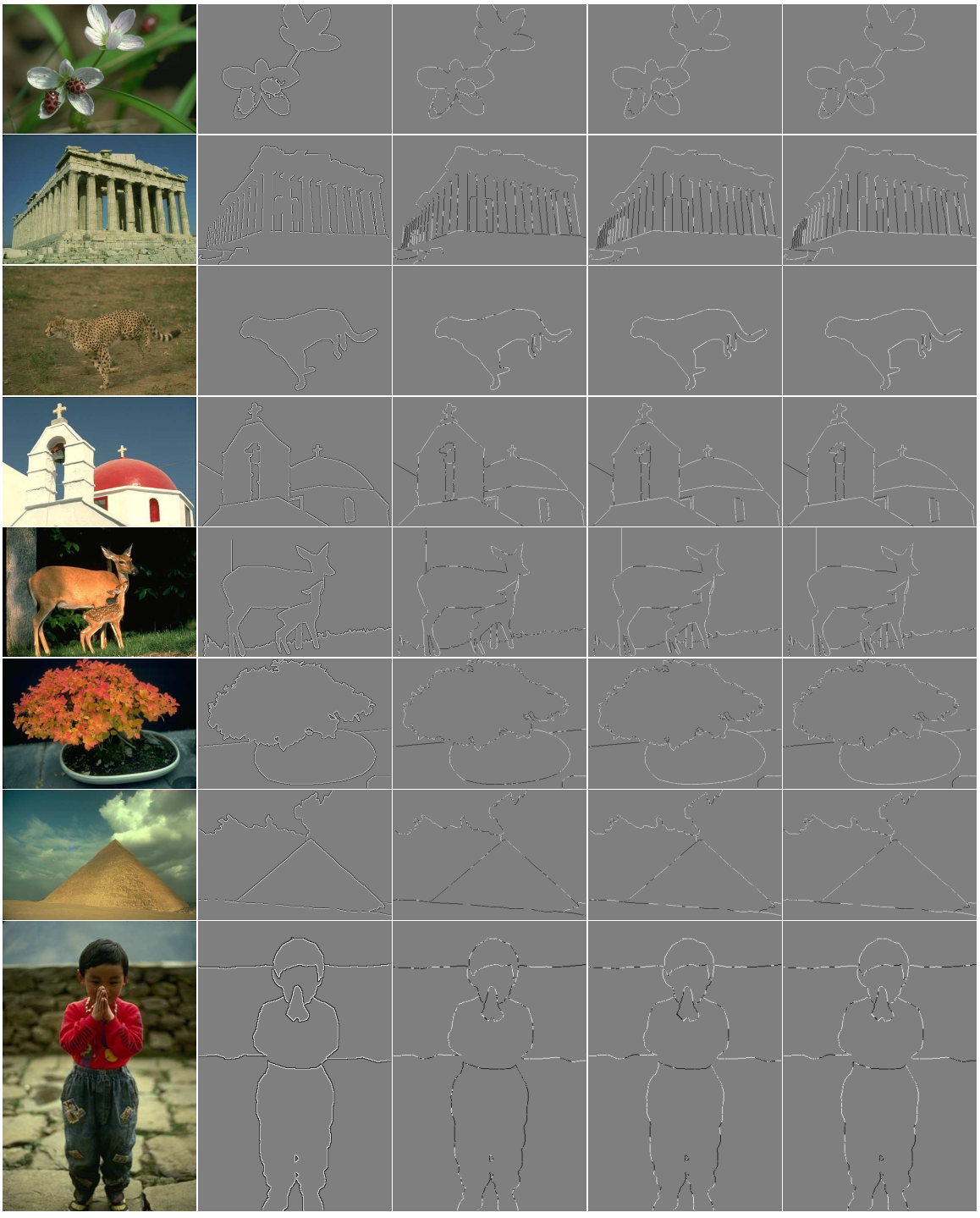

Figure 5.4: A few more examples of figure/ground classification results. The different columns of images here are arranged in the same order as in Figure 5.3

troduced into the model, which only influence the Orientation channel among the three feature channels. First, we show that even the reference model, with only a few global cues, convexity, surroundedness and parallelism, completely devoid of any local cues performs significantly better than chance level (50%) achieving a FGCA of 58.44% on the BSDS figure-ground dataset. Each local cue, when added alone leads to statistically significant improvement in the overall FGCA, compared to the reference model devoid of local cues, indicating their usefulness as independent local cues of FGO. The model with both SA and the T-Junctions, where T-Junctions are derived from automatically extracted edges, achieves an 8.77% improvement in terms of FGCA compared to that of the model without any local cues. When the T-Junctions obtained through automatically segmented edges are replaced by those derived from human drawn contours around objects, the improvement in FGCA of the model with both local cues increases to 12.04%. Moreover, the FGCA of the model with both local cues is always higher than that of the models with individual local cues, indicating the mutually facilitatory nature of local cues. In conclusion, SA and T-Junctions are useful, mutually beneficiary local cues and lead to statistically significant improvement in the FGCA of the feed forward, biologically motivated FGO model, either when added alone or together. In future, color opponency based local cues and global cues such as symmetry will be incorporated into the model.

Chapter 6

Modeling the influence of co-spatial audio on saliency of visual proto-objects

6.1 Introduction

Natural environment and our interaction with it is essentially multisensory, where we may deploy visual, tactile and/or auditory senses to perceive, learn and interact with our environment. In this part of the thesis (Chapters 6 and 7), our focus is on two important sensory modalities, vision, audition and their interaction. Perception in one sensory domain getting altered due to the influence of another sensory domain is a well studied phenomenon in psychophysics. Some famous examples are the McGurk effect [197], Ventriloquism [198], sound induced flash illusion [199], *etc.* For such altered perception in one sensory modality

due to the influence of some other sensory modality, there has to be some anatomical connections between those sensory modalities. In fact, there is enough experimental evidence now suggesting an interplay of connections (see Chapter 1, Section 1.1.3 for a detailed review of cross-modal anatomical pathways) between thalamus, primary sensory areas and higher level association areas [61, 69, 200].

The research related to how different senses influence each other is called by various terms such as cross-modal processing, multisensory integration, audio-visual integration (AVI), *etc.* In the context of this chapter, AVI is most relevant. AVI is predominantly observed in the Superior Colliculus area of cats in addition to many cortical and subcortical areas. One of the first neurophysiological experiments in cats [64] and subsequent ones [63, 65] have shown the following properties of audio-visual integration[1]:

- *Co-spatiality:* Audio-visual integration is stronger [64] when the constituent visual and auditory counterparts are spatially close to each other
- *Co-temporality:* when the visual and auditory counterparts are closer to each other in time, the effect of audio-visual integration is stronger [65]
- *Inverse effectiveness:* A phenomenon because of which the strength of multi-sensory excitation varies inversely with the strength of the most effective constituent unisensory stimulus strength [61, 83].

The phenomenon of inverse effectiveness is closely related to *super-additivity* observed in the multi-sensory neurons where the response of the multisensory neuron is much higher than

[1]Around the same time integration of audio-visual information in bimodal SC neurons was also observed in guinea-pigs by King and Palmer [201]

the simple sum of the responses evoked by unisensory constituent stimuli, when presented alone. This is typically observed with weak, near-threshold stimuli.

In this chapter, we investigate the nature of multisensory interaction between the auditory and visual domains. More specifically, we consider the effect of a spatially co-occurring auditory stimulus on the salience of a weak/inconspicuous visual target at the same spatial location among highly conspicuous visual distractors. Temporal concurrency is assumed between visual and auditory events. The motivation for this work is that audio-visual integration is highly effective when cue reliability is highly degraded in respective unisensory modalities. In such a scenario it is beneficial to integrate information from both sensory modalities in order to harness the advantages of each.

In the computational model that we present, auditory stimulus is modeled as a 1D Gaussian window centered close to the visually inconspicuous target. Visual saliency is computed using the proto-object based model of Russell et al. [25]. When auditory input is combined into the computation of visual saliency, our results show, the salience of the less conspicuous visual target increases, the behavioral manifestation of which could be increased accuracy in identifying the target and/or faster reaction time.

6.2 Related Work

The study of multi-sensory integration [60, 69, 200], specifically audio-visual integration [63, 202] has been an active area of research in neuroscience, psychology and cognitive science. In the computer science and engineering fields, there is an increased interest in the recent times [94, 203, 204]. For a detailed review of neuroscience and psychophysics

research related to audio-visual interaction, see [69, 200]. Here, we restrict our review to models of perception in audio-visual environments and some application oriented research using audio and video information. As the computational models of audio-visual integration and audio-visual saliency are limited, we combine the literature review for this chapter and Chapter 7 into one, which is covered here.

In one of the earliest works [205], a one-dimensional computational neural model of saccadic eye movement control by Superior Colliculus (SC) is investigated. The model can generate three different types of saccades: visual, multimodal and planned. It takes into account different coordinate transformations between retinotopic and head-centered coordinate systems, and the model is able to elicit multimodal enhancement and depression that is typically observed in SC neurons [64, 65]. However, the main focus is on Superior Colliculus function rather than studying audio-visual interaction from a salience perspective. A detailed model of the SC is presented in [206] with the aim of localizing audio-visual stimuli in real time. The model consists of 12,240 topographically organized neurons, which are hierarchically arranged into 9 feature maps. The receptive field of these neurons, which are fully connected to their input, are obtained through competitive learning. Intra-aural level differences are used to model auditory localization, while simple spatial and temporal differencing is used to model visual activity. A spiking neuron model [207] of audio-visual integration in barn owl uses Spike Timing Dependent Plasticity (STDP) to modulate activity dependent axon development, which is responsible for aligning visual and auditory localization maps. A neuromorphic implementation of the same using digital and analog mixed Very Large Scale Integration (mixed VLSI) can be found in [208].

In another neural model [209, 210] the visual and auditory neural inputs to the deep SC neuron are modeled as Poisson random variables. Their hypothesis is that the response of SC neurons is proportional to the presence of an audio-visual object/event in that spatial location which is conveyed to topographically arranged deep SC neurons via auditory and visual modalities. The model is able to elicit all properties of the SC neurons. An information theoretic explanation of super-additivity and other phenomena is given in a [210]. They also show that addition of a cue from another sensory modality increases the certainty of a target's location only if the input from initial modality/ies cannot reduce the uncertainty about target. Similar models are proposed in [211] and in [212], where the problem is formulated based on Bayes likelihood ratio. An important work [213] based on Bayesian inference explains a variety of cue combination phenomena including audio-visual spatial location estimation. According to the model, neuronal populations encode stimulus information using probabilistic population codes (PPCs) which represent probability distributions of stimulus properties of any arbitrary distribution and shape. They argue that neural populations approximate the Bayes rule using simple linear combination of neuronal population activities.

In [214], audiovisual arrays for untethered spoken interfaces are developed. The arrays localize the direction and distance of an auditory source from the microphone array, visually localize the auditory source, and then direct the microphone beamformer to track the speaker audio-visually. The method is robust to varying illumination and reverberation, and the authors report increased speech recognition accuracy using the AV array compared to non-array based processing.

In [215] the authors found that emotional saliency conveyed through audio, drags an observer's attention to the corresponding visual object, hence people often fail to notice any visual artifacts present in the video, suggest to exploit this property in intelligent video compression. For the same goal authors of [216] implement an efficient video coding algorithm based on the audio-visual focus of attention where sound source is identified from the correlation between audio and visual motion information. The same premise that audio-visual events draw an observer's attention is the basis for their formulation. A similar approach is applied to High Definition video compression in [217]. In these studies, spatial direction of sound was not considered, instead stereo or mono audio track accompanying the video was used in all computational and experimental work.

In [218], a multimodal bottom-up attentional system consisting of a combined audio-visual salience map and selective attention mechanism is implemented for the humanoid robot iCub. The visual salience map is computed from color, intensity, orientation and motion maps. The auditory salience map consists of the location of the sound source. Both are registered in ego-centric coordinates. The audio-visual salience map is constructed by performing a pointwise *max* operation on visual and auditory maps. In an extension to multi-camera setting [219], the 2D saliency maps are projected into a 3D space using ray tracing and combined as a fuzzy aggregations of salience spaces. In [220, 221], after computing the audio and visual saliency maps, each salient event/proto-object is parameterized by salience value, cluster center (mean location), and covariance matrix (uncertainty in estimating location). The maps are linearly combined based on [222]. Extensions of this approach can be found in [223]. A work related to [223] is presented in [224] where weighted

linear combination of proto-object representations obtained using mean-shift clustering is detailed. Even though the method uses linear combination, the authors do not use motion information in computing the visual saliency map. A Self Organizing Map (SOM) based model of audio-visual integration was presented in [225] in which the transformations between sensory modalities, and the respective sensory reliabilities are learned in an unsupervisory manner. A system to detect and track a speaker using a multi-modal, audio-visual sensor set that fuses visual and auditory evidence about the presence of a speaker using Bayes network was presented in [226]. In a series of papers [203, 227, 228] audio-visual saliency is computed as a linear mixture of visual and auditory saliency maps for the purpose of movie summarization and key frame detection. No spatial information about audio is considered. The algorithm performs well in summarizing the videos for informativeness and enjoyability for movie clips of various genres. An extension of these models incorporating text Saliency can be found in [229]. By assuming a single moving sound source in the scene, audio was incorporated into the visual saliency map in [230] where sound location was associated with the visual object by correlating sound properties with the motion signal. By computing Bayesian surprise as in [231], the authors in [232] present a visual attention model driven by auditory cues, where surprising auditory events are used to select synchronized visual features and emphasize them in a audio-visual surprise map. A real-time multi-modal home entertainment system [233] performing a Just-In-Time association of features related to a person from audio and video are fused based on the shortest distance between each of the faces (in video) and the audio direction vector. In an intuitive study [234] speaker localization by measuring the audio-visual synchrony in terms of mutual

information between auditory features and pixel intensity change is considered. In a single active speaker scenario, they obtain good preliminary results. No microphone arrays are used for the localization task. In [235] visually detected face location is used to improve the speaker localization using a microphone array. A fast audiovisual attention model for human detection and localization is proposed in [236].

The effect of sound on gaze behavior in videos was studied in [204, 237] where a preliminary computational model to predict eye movements was proposed. They use motion information to detect sound source. High level features such as face are hand labeled. A comparison of eye movements during visual only and audio-visual conditions with their model shows that adding sound information improved predictive power of their model. The role of salience, faces and sound in directing the attention of human observers (measured by gaze tracking) was studied with psychophysics experiments and computational modeling in [238] and an audiovisual attention model for natural conversation scenes was proposed in [239], where they use a speaker diarization algorithm to compute saliency[2]. Even though their study is restricted to conversation of humans and not applicable any generic audio-visual scene, hence cannot be regarded as a generalized model of audio-visual saliency, some interesting results are shown. Using EM algorithm to determine the individual contributions of bottom-up salience, faces and sound in gaze prediction, they show adding original speech to video improves gaze predictability, whereas adding irrelevant speech or unrelated natural sounds has no effect. By using speaker diarization algorithm [239] when the weight

[2]Speaker diarization deals with the segmentation of speech into non-overlapping homogeneous regions separated by silence and assigning each of the segmented speech bits to unique speakers. Even though multimodal speaker diarization methods perform audio-visual integration, for example in [240] they combine audio and visual information using an Expectation Maximization (EM) algorithm in a Dynamic Bayesian Network (DBN) framework, the application is specific to speech and cannot generalize to a saliency map, hence not reviewed. A review of these methods can be found in [241].

for active speakers was increased, their audio-visual attention model significantly outperformed the visual saliency model with equal weights for all faces. An audio-visual saliency map is developed in [242] where features such as color, intensity, orientation, faces, speech are linearly combined with unequal weights to give different types of saliency maps depending on the presence/absence of faces and/or speech. It is not clear as to whether location of the sound was used in their approach. Plus, they do not factor in motion, which is an important feature while designing a saliency map for moving pictures.

6.3 Data and Methods

The effectiveness of audio-visual integration in detecting weakly visible visual target among many highly conspicuous distractors is studied by computing an audio-visual (AV) integration map. It should be noted that there are many ways in which audio-visual integration (AVI) can be modeled and what we compute here is one such implementation, which we call the AV integration map. The visual stimuli (target and distractors) are deliberately made barely distinguishable from each other. If the auditory stimulus helps identify the target, then the AVI map should reflect the same result. The effectiveness of AVI with respect to unimodal saliency maps (SM) is studied for different stimulus conditions.

6.3.1 Visual Stimuli

The visual stimuli are rectangular images with a width of 1800 pixels and height of 150 pixels (Figure 6.1). A horizontal reference line guides the observer to possible locations of the target and distractors. The task is to identify a weakly visible target symbol among

Figure 6.1: Visual stimulus with target ('Y') and distractors ('X'). The distractors are visually more conspicuous than target

a number of more conspicuous visual distractors in the audio-visual scene. The targets are displayed as the letter 'Y' and distractors are displayed as the letter 'X' (Figure 6.1). The number of visual distractors, D_V, is randomly chosen to be between 1 and 5 inclusive. There is always *only* one target in every stimulus image. Neither the target nor distractors are allowed to lie within 10 pixels from the image boundaries to avoid unwanted artifacts from the visual salience computation. Distractor locations are randomly selected without replacement from all possible spatial locations on the abscissa. Among the remaining locations, a target location is randomly chosen. Care is taken to avoid symbols flanking too close to each other. The intensities of both target and distractors are kept identical to avoid intensity-related salience differences. Salience differences in our stimuli are observed because of differences in shape of the symbols only.

Both the distractors and target are distinguishable from the background, but identifying the target from the distractors is a difficult task. If we rely on using the visual domain alone to locate the target, this search requires a considerable amount of attention and thus serial processing to identify if each symbol is the target.

6.3.2 Auditory Stimuli

The auditory space is modeled to be spatially coincident with the visual space covering the entire image. We simulate the activation of one of the 8 speakers that are placed

equidistant to each other covering the visual space of the entire image. So, the auditory space is divided into 8 equal sections. If the visual target ('Y') is present in a specific section, a Gaussian window with zero mean and unit variance is centered in that section to represent the approximate auditory signal location. Since auditory localization is generally less precise than visual localization we center the envelope in a particular section of the map irrespective of the exact location of the visual target within that section.

Our model for the auditory signal also serves as an auditory salience map (ASM) because we take spatial location of sound stimulus to be the only relevant feature. Hence, the ASM consists of an activation region if a sound stimulus originates from that location. The sound localization inaccuracy observed in both humans and primates is the motivation to model the stimulus as a Gaussian window (Eq. 6.1) situated at the location of the sound stimulus:

$$A(x) = \begin{cases} e^{-\frac{1}{2}\left(\alpha \frac{(x-\frac{x_o}{2})}{\frac{x_o}{2}}\right)^2} & \text{if } x \in Q_v \\ \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

In Eq. 6.1, Q_v represents the section in which the visual target lies, x_o is the width of the window equal to the length of the active section. The parameter $\alpha = 2.5$, reciprocal of the standard deviation controls the width of the window [111]. The width of the Gaussian roughly corresponds to an uncertainty in auditory stimulus location and the height corresponds to the reliability of the auditory cue.

6.3.3 Audio-Visual Integration map

We first compute a proto-object based salience map [25] of the visual scene to investigate the relative visual salience of target and distractors. In the auditory domain, since stimulus location is the only feature considered, the stimulus location map (Figure 6.3) also serves as the auditory saliency map which is already computed. The visual and auditory saliency maps are combined multiplicatively as:

$$\mathcal{M}_{AVI} = f(A) \otimes V \tag{6.2}$$

$$= (1 + A) \otimes V, \tag{6.3}$$

$$\text{where } V = \frac{1}{3}(\mathcal{N}_2(\bar{\mathcal{O}}) + \mathcal{N}_2(\bar{\mathcal{I}}) + \mathcal{N}_2(\bar{\mathcal{C}})). \tag{6.4}$$

In Eqs. 6.2-6.4, A is the auditory salience/location map, $\mathcal{M}_{AVI}$ is the audio-visual integration map and V is the proto-object based visual saliency map. The normalization operator is denoted by $\mathcal{N}_2(\cdot)$ which is the same as the one used in Itti et al. [21], which accentuates strong isolated activity and suppresses many weak activities, and point-wise multiplication is denoted by the symbol $\otimes$. Color, orientation and intensity conspicuity maps are denoted by $\bar{\mathcal{C}}$, $\bar{\mathcal{O}}$ and $\bar{\mathcal{I}}$, respectively. For more details of visual proto-object based saliency computation, please refer to [25]. By combining the auditory and visual saliency maps as shown in Eq. 6.2, we retain all salient visual stimuli and also enhance the salience

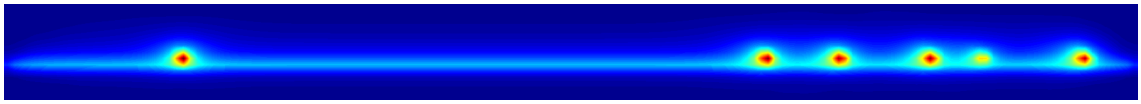

Figure 6.2: Proto-object saliency map of the visual stimulus. Notice that target is less salient than distractors

Figure 6.3: Auditory stimulus which is also the ASM modeled as a one-dimensional Gaussian. Width of the Gaussian corresponds to uncertainty in location, height to signal reliability

of only those visual stimuli that have a spatially co-occurring salient event in the auditory domain.

6.4 Results and Discussion

The visual proto-object based salience map is computed with default parameters listed in [25]. In the visual domain (Figure 6.2) we see that distractors are more salient than the target. This salience result implies that an observer is more likely to shift his or her attention to the distractors than to the target. In such a scenario, identifying the target requires an elaborate visual search. On the other hand (see Figure 6.3), in the auditory domain the section in which the target lies is salient, but the exact location of visual stimulus cannot be identified.

We model the integration of visual and auditory saliencies in a combined audio-visual integration map as described in Eq. 6.2. The combined AVI map is shown in Figure 6.4.

The combined AV integration map illustrates the idea that combining information maps

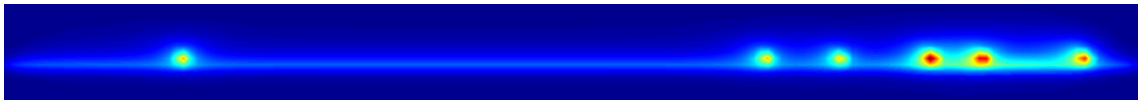

Figure 6.4: Combined audio-visual integration map. Notice the enhancement of saliency of the target

from multiple modalities can aid the search for the less conspicuous target in an environment among more prominent visual distractors. Even though the visual salience map shows the target as less conspicuous than the distractors, adding the auditory information allows for approximate identification of the location of the target. Without the auditory stimulus, visual distractors exhibiting higher salience than the target are attended to first. However, when an auditory stimulus co-occurs with the target location, the less conspicuous target becomes more salient than the visual distractors due to multisensory interaction between the auditory and visual modalities. Another example of audio-visual integration map for a different stimulus condition is shown in Figure 6.5.

Our results confirm the effectiveness of audio-visual integration when cues in unisensory modalities are weak, and therefore cannot elicit a strong response based on unisensory cue alone. The effectiveness of multisensory integration is inversely related to effectiveness of unimodal cues [63]. Since we observe increased multisensory salience for the weakly visible target, our model exhibits a form of inverse effectiveness as reported in previous studies [243].

Our model can be advantageous compared to that of [218] because the latter model only highlights salient regions from individual domains. For example, in a scenario where there are three types of events (unimodal auditory, unimodal visual and bimodal audiovisual), the audiovisual event should be more salient than the unimodal events. However, the

model from [218], which computes audio-visual saliency by taking maximum value across all channels at each location, will not be able to assign a higher salience to the bimodal event compared to unimodal events. On the other hand, our model assigns higher salience to bimodal events as compared to unimodal ones. Our model also agrees with previous studies [222, 244] where lateralized auditory stimulation was found to topographically increase the salience of the visual field. The model favorably compares with some other experiments where stimulus conditions are slightly different, but visual response enhancement was observed. In [92], a sudden sound, spatially coincident with a *subsequently* occurring visual stimulus was found to improve the detectability of the flash. Our model shows evidence for their main conclusion that involuntary attention to spatially registered sound enhances visual response. In [245] event related potentials were recorded in an audio-visual integration experiment where they found addition of task irrelevant auditory stimulus increased the accuracy and decreased the reaction time in correctly identifying a visual target. It is in agreement with our model. We only consider spatial overlap of auditory and visual stimuli; the temporal aspects have not been accounted for, which is a limitation of our model.

6.5 Conclusion

We present a way of combining separate auditory and visual salience maps into an audio-visual integration map where saliency maps from their respective modalities are combined multiplicatively. We retain saliencies of all visual stimuli while enhancing the salience of the target visual stimulus in a model of audio-visual interaction. Without the auditory stimulus, the visual distractors exhibit higher salience compared to the visual target. However,

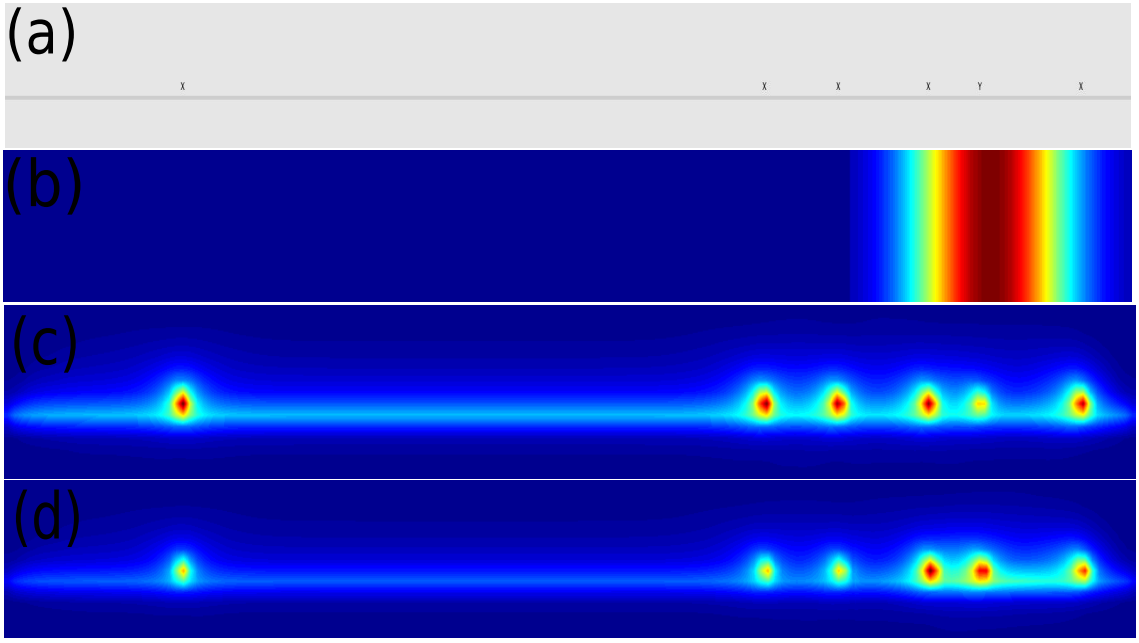

Figure 6.5: Another example of AV integration: (a) Visual stimuli; (b) Auditory Saliency Map; (c) Visual proto-object based saliency map; (d) Audio-visual integration map. Notice the target is more conspicuous compared to the distractors, a reversal effect.

when an auditory stimulus co-occurs with the target visual location the effect is reversed, making the visual target more salient than the distractors. Our results agree with previous neurophysiological studies [243] which establish that audio-visual integration is highly effective when the cue reliability is low in individual modalities taken separately.

Chapter 7

A proto-object based audiovisual saliency map

7.1 Overview

Scientists and engineers have traditionally separated the analysis of a multisensory scene into its constituent sensory domains. In this approach, for example, all auditory events are processed separately and independently of visual and/or somatosensory streams even though the same multisensory event might have created those constituent streams. It was previously necessary to compartmentalize the analysis because of the sheer enormity of information as well as the limitations of experimental techniques and computational resources. With recent advances, it is now possible to perform integrated analysis of sensory systems including interactions within and across sensory modalities. Such efforts are becoming increasingly common in cellular neurophysiology, imaging and psychophysics studies [69, 83, 198, 202,

243].

Recent evidence from neuroscience [60, 83] suggests that the traditional view that the low level areas of cortex are strictly unisensory, processing sensory information independently, which is later on merged in higher level associative areas is increasingly becoming obsolete. This has been proved by many fMRI [84, 85], EEG [86] and neuro-physiological experiments [87, 88] at various neural population scales. There is now enough evidence to suggest an interplay of connections between thalamus, primary sensory and higher level association areas which are responsible for audiovisual integration. The broader implications of these biological findings may be that learning, memory and intelligence are tightly associated with the multi-sensory nature of the world.

Hence, incorporating this knowledge in computational algorithms can lead to better scene understanding and object recognition for which there is a great need. Moreover, combining visual and auditory information to associate visual objects with their sounds can lead to better understanding of events. For example, discerning whether the bat hit the baseball during a swing of the bat, tracking objects under severe occlusions, poor lighting conditions *etc* can be more accurately performed only when we take audio and visual counterparts together. The applications of such technologies are numerous and in varied fields.

In summary, a better understanding of interaction, information integration, and complementarity of information across senses may help us build many intelligent algorithms for scene analysis, object detection and recognition, human activity and gait detection, elder/child care and monitoring, surveillance, robotic navigation, biometrics *etc,* with better

performance, stability and robustness to noise. In one application, for example, fusing auditory (voice) and visual (face) features improved the performance of speaker identification and face recognition systems [246, 247]. Hence, our objective in this chapter is to develop a scene analysis algorithm using multisensory information, specifically vision and audio. We develop a purely bottom-up, proto-object based audiovisual saliency map (AVSM) for the analysis of dynamic natural scenes.

Kimchi et al. [248] show that objects attract human attention in a bottom-up manner. Also, Nuthmann and Henderson [249] find that eye fixations of human observers tend to coincide with object centers. Simple pixel based saliency models determine salient locations based on dissimilarities in low-level features such as color, orientation, *etc*, which do not always coincide with object centers. In a proto-object based saliency model, the grouping mechanism integrates border ownership (Section 7.2.3) information in an annular fashion to create selectivity for objects. This is based on Gestalt properties of convexity, surroundedness and parallelism. Hence the predicted salient locations in our model generally coincide with object centers, which cannot be achieved with simple feature based saliency models.

Building on the work of Russell et al. [25], we add visual motion (Section 7.2.1.1) as another independent feature type along with Color, Intensity and Orientation in the visual domain, all of which undergo a grouping process (Section 7.2.4) to form proto-objects of each feature type. In the auditory domain, we consider the location and intensity of sound as the only proto-objects as these are found to be most influential in drawing the spatial attention of an observer in many psycho-physics studies. Various methods of combination of the auditory and visual proto-object features are considered (Section 7.2.6). We demonstrate

the efficacy of the AVSM in predicting salient locations in the audiovisual environments by testing it on real world AV data collected from a specialized hardware (Section 7.3) that can collect 360^0 audio and video that are temporally synchronized and spatially co-registered. The AVSM captures nearly all visual, auditory and audio-visually salient events, just as any human observer would notice in that environment[1].

7.2 Description of the model

The computation of audiovisual saliency map is similar to the computation of proto-object based visual saliency map for static images explained in [25], except for (i) the addition of two new feature channels, the visual motion channel and the auditory loudness and location channel; and (ii) different ways of combining the conspicuity maps to get the final saliency maps. Hence, whereever the computation is identical to [25], we will only give a gist of that computation to avoid repetition and detailed explanation otherwise.

The AVSM is computed by grouping auditory and visual bottom-up features at various scales, then normalizing the grouped features within and across scales, followed by merging features across scales and linear combination of the resulting feature conspicuity maps (Figure 7.1). The features are derived from the color video and multi-channel audio input (Section 7.3) without any top-down attentional biases, hence the computation is purely bottom-up. And the mechanism of grouping "binds" features within a channel into candidate objects or "proto-objects". Approximate size and location are the only properties of objects that the grouping mechanism estimates, hence they are termed "proto-objects".

[1]The literature review related to audio-visual saliency was covered in Chapter 6, Section 6.2

Such proto-objects, many of them, form simultaneously and dissolve rapidly [250] in a purely bottom-up manner. Top-down attention is required to hold them together into coherent objects.

Also, computation of AVSM is completely feed-forward. Many spatial scales are used to achieve scale invariance. First the independent feature maps are computed, features within each channel are grouped into proto-objects. Such proto-object feature pyramids at various scales are normalized within and across scales. Then they are merged across scales followed by normalization across feature channels to give rise to conspicuity maps. The conspicuity maps are combined in different ways to get the three types of AVSMs. Each of these steps is explained in more detail below.

7.2.1 Computation of feature channels

We consider color, intensity, orientation and motion as separate, independent feature channels in the visual domain. Loudness and spatial location as features in the auditory space. The audio-visual camera equipment used for data gathering (See Section 7.3) guarantees spatial and temporal concurrency of audio and video.

A single intensity channel, where intensity is computed as the average of Red, Green and Blue color channels is used. Four feature sub-channels for angle, $\theta = \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$ are used for Orientation channel. Four color opponency feature sub-channels: Red-Green ($\mathcal{RG}$), Green-Red ($\mathcal{GR}$), Blue-Yellow ($\mathcal{BY}$) and Yellow-Blue ($\mathcal{YB}$) are used for color channel. The computation of color, orientation and intensity feature channels is identical to Russell et al. [25].

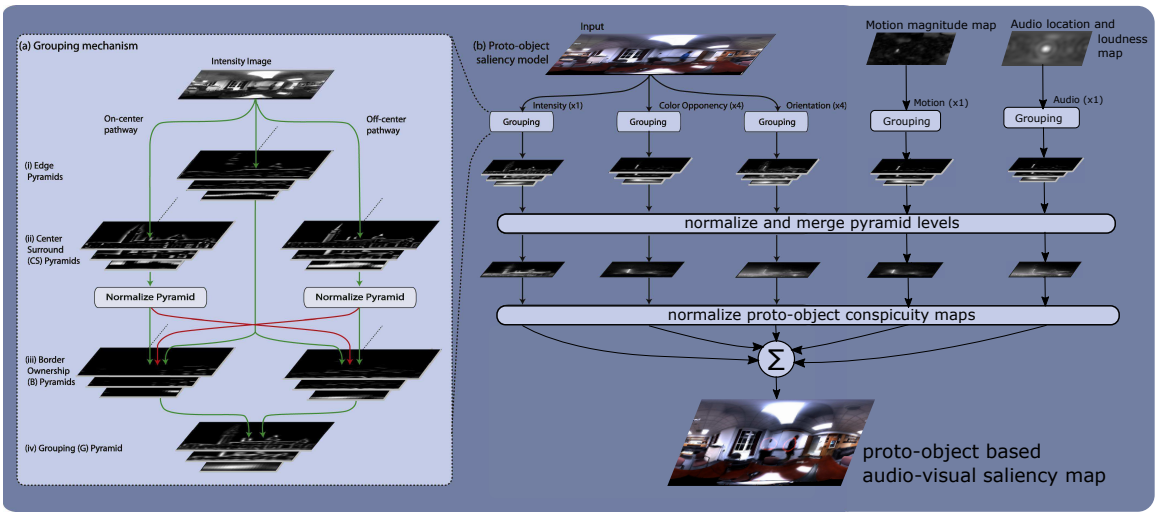

Figure 7.1: Computation of proto-object based AVSM. (a) Grouping mechanism: At each scale, feature maps are filtered with center-surround cells, normalized, followed by border ownership (BO) computation. Grouping cells, at each scale, receive a higher feed-forward input from BO cells if they are consistent with Gestalt properties of convexity, proximity and surroundedness. Hence, grouping gathers conspicuous BO activity of features at object centers. Each feature channel undergoes same computation at various scales (except Orientation, see explanation), only Intensity channel shown in Figure 7.1(a). (b) Audivisual saliency computation mechanism: Five feature types are considered: Color, Orientation, Intensity and Motion in the visual domain; spatial location and loudness estimate in the auditory domain. All features undergo grouping as explained in (a) to obtain proto-object pyramids. The proto-object pyramids are normalized, collapsed to get the feature specific conspicuity maps such that isolated strong activity is accentuated and distributed weak activity is suppressed. The conspicuity maps are then combined in different ways as explained in Section 7.2.6 to get different saliency maps. Adapted from Figure. 5 of [25] with permission.

Visual motion, auditory loudness and location estimate are the newly added features, the computation of which is explained below.

7.2.1.1 Visual motion channel

Motion is computed using the optical flow algorithm described in [251] and the corresponding code available at [252]. Consider two successive video frames, $I(x, y, t)$ and $I(x, y, t + 1)$. If the underlying object has moved between t and $t + 1$, then the intensity at pixel location, (x, y) at time, t should be the same in a nearby pixel location at $(x + \Delta x, y + \Delta y)$ in the successive frame at $t + 1$. Using this as one of the constraints, the flow is estimated which gives the horizontal and vertical velocity components, $u(x, y, t)$ and $v(x, y, t)$ respectively, at each pixel location (x, y) at time t. For a more detailed explanation, see [251]. Since we are interested in detecting salient events only, we do not take into account the exact motion at each location as given by $u(x, y, t)$ and $v(x, y, t)$, instead, we look at how big the motion is at each location in the image. The magnitude of motion at each location is computed as,

$$M(x, y, t) = \sqrt{u(x, y, t)^2 + v(x, y, t)^2} \tag{7.1}$$

The motion map, $M(x, y, t)$ gives the magnitude of motion at each location in the visual scene at different time instances, t. Figure 7.2 shows two successive frames of the video in (A) and (B) respectively. The computed magnitude of motion[2] using the optic flow method in [251] is shown in (C). The person in the video is the only moving source.

[2]for an alternate method to compute motion and incorporate it into saliency map, see [158]

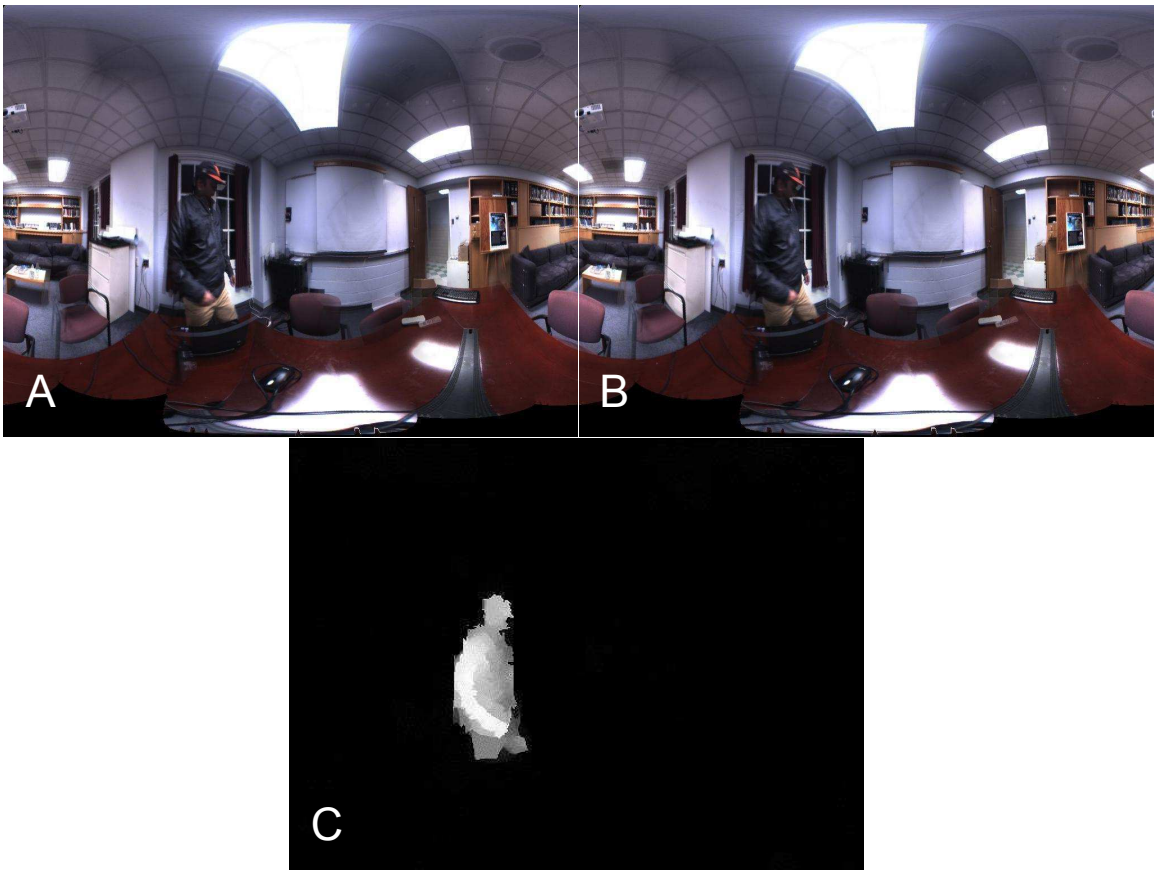

Figure 7.2: Computation of motion magnitude map. (A) A frame from the dataset, (B) the successive frame. The person in the two frames is the only moving object in the two image frames, moving from left to right in the image. (C) The motion magnitude map

7.2.1.2 Auditory loudness and location channel

The auditory input consists of a recording of the 3D sound field using 64 microphones arranged on a sphere (See Section 7.3 for details). We compute a single map for both loudness and location of sound sources, $A(x, y, t)$. The value at a location in the map, $A(x, y, t)$ gives an estimate of loudness at that location at time, t, hence we simultaneously get an estimate of the presence and loudness of sound sources at every location in the entire environment. These two features are computed using beamforming technique as described in [253, 254, 255]. A more detailed account is given in Section 7.3. Two different frames of video and the corresponding auditory loudness and location maps superimposed on the visual images are shown in Figure 7.3 (B) and (F) respectively. Warm colors indicate higher intensity of sound from that location in the video.

7.2.2 Feature pyramid decomposition

Feature pyramids are computed for each type. As the scale increases, the resolution of the feature map decreases. The feature maps of successively higher scales are computed by downsampling the feature map from the previous scale. The downsampling factor can be either $\sqrt{2}$ (half-octave) or 2 (full octave). The feature pyramids thus obtained are used to compute proto-objects by border ownership and grouping computation process explained the next two sections.

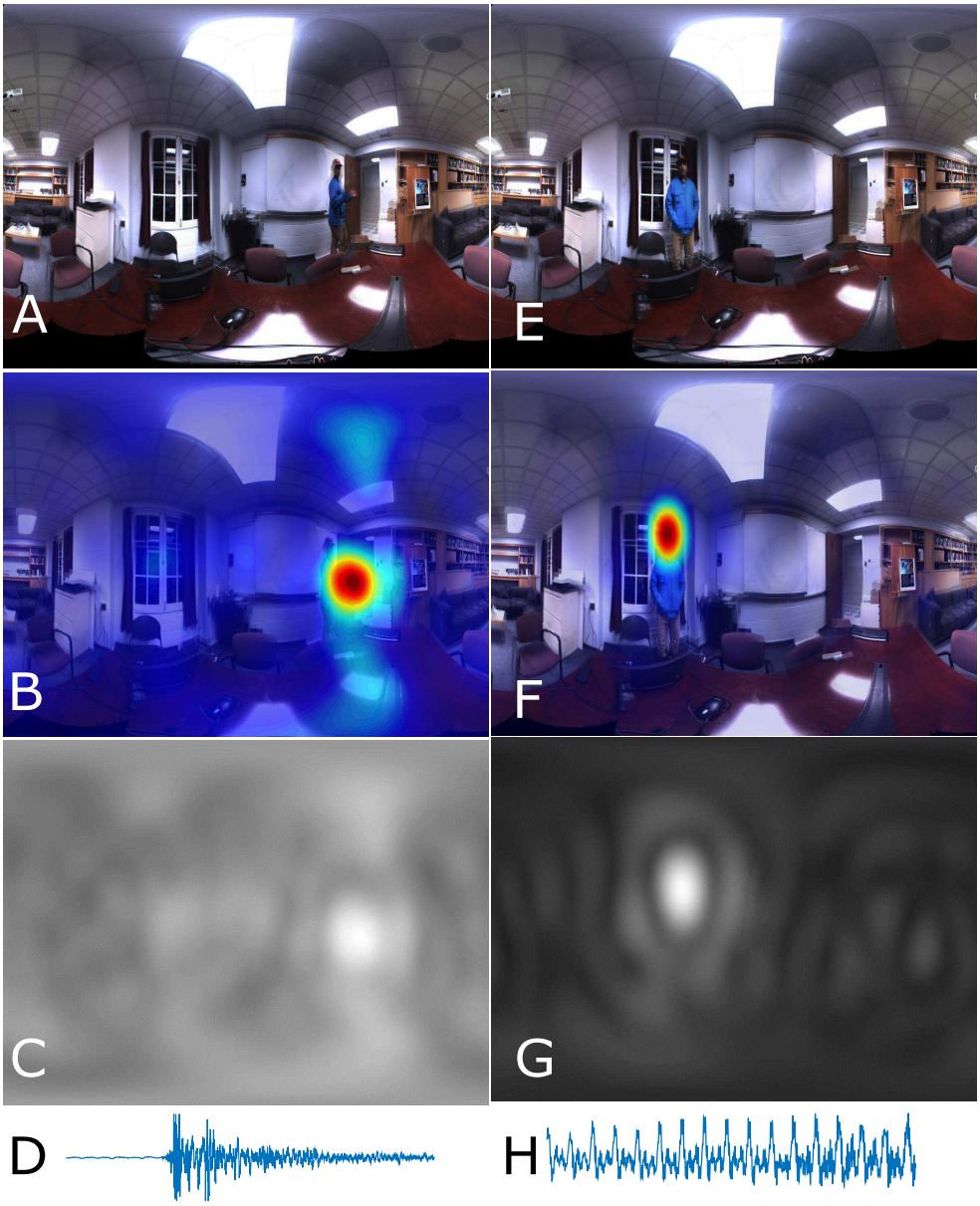

Figure 7.3: Computation of auditory location and loudness map: Two different video frames in (A) and (E). Corresponding audio samples from one of the audio channels are shown in (D) and (H), which display only 4410 samples, *i.e.* , audio of 0.1 s. Corresponding auditory loudness and location maps are in (C) and (G). Corresponding video overlays in (B) and (F) for illustration only, not used in any computation. Left column: sound of a clap. Right column: some part of the word "eight" uttered by the person

7.2.3 Border ownership pyramid computation

Computation of proto-objects by grouping mechanism can be divided into two sub-steps: (i) border ownership pyramid computation, and (ii) grouping pyramid computation.

The operations performed on any of the features, auditory or visual is the same. Edges of four orientations, $\theta = \{0, \pi/4, \pi/2, 3\pi/4\}$ are computed using the Gabor filter bank. The V1 complex cell responses [25] thus obtained are used to construct the edge pyramids. Border ownership response is computed by modulating the edge pyramid by the activity of center-surround feature differences on either side of the border. The rationale behind this is the observation made by Zhang and von der Heydt [256], where they reported that the activity border ownership cells was enhanced when image fragments were placed on their preferred side, but suppressed for the non-preferred side.

Two types of center-surround (CS) feature pyramids are used. The center-surround light pyramid detects strong features surrounded by weak ones. Similarly, to detect weak features surrounded by a strong background, a center-surround dark pyramid is used. The rationale behind light and dark CS pyramids is that there can be bright objects on a dark background and *vice-versa*. The center-surround pyramids are constructed by convolving feature maps with Difference of Gaussian (DoG) filters.

The CS pyramid computation is performed in this manner for all feature types including motion and audio, except for the orientation channel. For the orientation feature channel, the DoG filters are replaced by the even symmetric Gabor filters which detect edges. This is because, for the orientation channel, feature contrasts are not typically symmetric as in the case of other channels, but oriented at a specific angle.

An important step in the border ownership computation is normalization of the center-surround feature pyramids. We follow the normalization method used in [21] which enhances isolated high activities and suppresses many closely clustered similar activities.

This normalization step enables comparison of light and dark CS pyramids. Because of the normalization, border ownership activity and grouping activity are proportionately modulated, deciding relative salience of proto-objects from the grouping activity.

The border ownership (BO) pyramids corresponding to light and dark CS pyramids are constructed by modulating the edge activity by the normalized CS pyramid activity. The light and dark BO pyramids are merged across scales and summed to get contrast polarity invariant BO pyramids. For each orientation, two BO pyramids with opposite BO preferences are be computed. From this, the winning BO pyramids are computed by a winner-take-all mechanism.

7.2.4 Grouping pyramid computation

The grouping computation shifts the BO activity from edge pixels to object centers. Grouping pyramids are computed by integrating the winning BO pyramid activity such that selectivity for Gestalt properties of convexity, proximity and surroundedness is enhanced. This is done by using Grouping cells in this computation, which have an annular receptive field. The shape of G cells gives rise to selectivity for convex, surrounded objects. At this stage we have the grouping or proto-object pyramids which are normalized and combined across scales to compute feature conspicuity maps, and then the saliency map.

7.2.5 Normalization and across-scale combination of grouping pyramids

The computation of grouping pyramids as explained in Section 7.2.4 is performed for each feature type. Let us represent the grouping pyramid for intensity feature channel by $\mathcal{G}^k_{\mathcal{I}}(x,y,t)$, where k denotes the scale of the proto-object map in the grouping pyramid. The color feature sub-channel grouping pyramids are represented as $\mathcal{G}^k_{\mathcal{RG}}(x,y,t)$ for Red-Green, $\mathcal{G}^k_{\mathcal{GR}}(x,y,t)$ for Green-Red, $\mathcal{G}^k_{\mathcal{BY}}(x,y,t)$ for Blue-Yellow and $\mathcal{G}^k_{\mathcal{YB}}(x,y,t)$ for Yellow-Blue color opponencies. The orientation grouping pyramids are denoted by $\mathcal{G}^k_{\mathcal{O}}(x,y,t,\theta)$ where θ denotes orientation, motion feature channel by $\mathcal{G}^k_{\mathcal{M}}(x,y,t)$ and auditory location and intensity feature channel by $\mathcal{G}^k_{\mathcal{A}}(x,y,t)$. The corresponding conspicuity maps for , intensity $\mathcal{I}(x,y,t)$, color $\mathcal{C}(x,y,t)$, orientation $\mathcal{O}(x,y,t)$, motion, $\mathcal{M}(x,y,t)$ and auditory location and loudness estimate, $\mathcal{A}(x,y,t)$ are respectively obtained as,

$$\mathcal{I}(x,y,t) = \bigoplus_{k=1}^{k=10} \mathcal{N}_2(\mathcal{G}^k_{\mathcal{I}}(x,y,t)) \tag{7.2}$$

$$\begin{aligned}\mathcal{C}(x,y,t) = \bigoplus_{k=1}^{k=10} \Big(&\mathcal{N}_2(\mathcal{G}^k_{\mathcal{RG}}(x,y,t)) + \mathcal{N}_2(\mathcal{G}^k_{\mathcal{GR}}(x,y,t)) \\ &+\mathcal{N}_2(\mathcal{G}^k_{\mathcal{BY}}(x,y,t)) + \mathcal{N}_2(\mathcal{G}^k_{\mathcal{YB}}(x,y,t))\Big)\end{aligned} \tag{7.3}$$

$$\mathcal{O}(x,y,t) = \bigoplus_{k=1}^{k=10} \sum_{\theta\in\{0,\frac{\pi}{4},\frac{\pi}{2},3\frac{\pi}{4}\}} \mathcal{N}_2(\mathcal{G}^k_{\mathcal{O}}(x,y,t,\theta)) \tag{7.4}$$

$$\mathcal{M}(x,y,t) = \bigoplus_{k=1}^{k=10} \mathcal{N}_2(\mathcal{G}^k_{\mathcal{M}}(x,y,t)) \tag{7.5}$$

$$\mathcal{A}(x,y,t) = \bigoplus_{k=1}^{k=10} \mathcal{N}_2(\mathcal{G}^k_{\mathcal{A}}(x,y,t)) \tag{7.6}$$

where $\mathcal{N}_2(\cdot)$ is a normalization step as explained in Itti et al. [21], which accentuates strong isolated activity and suppresses many weak activities, the symbol $\bigoplus$ denotes "across-scale" addition of the proto-object maps, which is done by resampling (up- or down-sampling depending on the scale, k) maps at each level to a common scale (in this case, the common scale is $k = 8$) and then doing pixel-by-pixel addition. We use the same set of parameters as in Table 1 of Russell et al. [25] for our computation as well.

The conspicuity maps, due to varied number of feature sub-channels have different ranges of activity, hence if we linearly combine without any rescaling to a common scale, those features with higher number of sub-channels may dominate. Hence, each feature conspicuity map is rescaled to the same range, $[0, \ldots, 1]$. The conspicuity maps are combined in different ways to get different types of saliency maps as explained in Section 7.2.6.

7.2.6 Combination of conspicuity maps

The visual saliency map is computed as,

$$\begin{aligned} \mathcal{VSM}(x,y,t) = w_I \mathcal{R}(\mathcal{I}(x,y,t)) + w_C \mathcal{R}(\mathcal{C}(x,y,t)) \\ + w_O \mathcal{R}(\mathcal{O}(x,y,t)) + w_M \mathcal{R}(\mathcal{M}(x,y,t)) \end{aligned} \tag{7.7}$$

where $\mathcal{VSM}(x, y, t)$ is the visual saliency map, $\mathcal{R}(.)$ is the rescaling operator that rescales each map to the same range, $[0, \ldots, 1]$ and w_I, w_C, w_O and w_M are the individual weights for intensity, color, orientation and motion conspicuity maps, respectively. In our implementation, all weights are equal and each is set to 0.25, *i.e.* , $w_I = w_C = w_O = w_M = \frac{1}{4}$.

Since audio is a single feature channel, the conspicuity map for auditory location and loudness is also the auditory saliency map, $\mathcal{ASM}(x, y, t)$.

We compute the audio-visual saliency map in three different ways to compare the most effective method to identify salient events (See Section 7.4 for related discussion).

In the first method a weighted combination of all feature conspicuity maps is done to get the audio-visual saliency map as,

$$\begin{aligned}\mathcal{AVSM}_1(x, y, t) = w_I\mathcal{R}(\mathcal{I}(x, y, t)) + w_C\mathcal{R}(\mathcal{C}(x, y, t)) + w_O\mathcal{R}(\mathcal{O}(x, y, t)) \\ + w_M\mathcal{R}(\mathcal{M}(x, y, t)) + w_A\mathcal{R}(\mathcal{A}(x, y, t))\end{aligned} \tag{7.8}$$

where different weights can be set for $w_{(.)}$ such that the sum of all weights equals 1. In our implementation, all weights are set equal, *i.e.* , $w_I = w_C = w_O = w_M = w_A = \frac{1}{5}$.

In the second method, the visual saliency map is computed as in Equation 7.7 and then a simple average of the visual saliency map and the auditory conspicuity map (also auditory saliency map, $\mathcal{ASM}(x, y, t)$) is computed to get the audio-visual saliency map as,

$$\mathcal{AVSM}_2(x, y, t) = \frac{1}{2}\Big(\mathcal{R}(\mathcal{VSM}(x, y, t)) + \mathcal{R}(\mathcal{A}(x, y, t))\Big) \tag{7.9}$$

The distribution of weights in Equation 7.9 is different from that in Equation 7.8. In method 2, a "late combination" of the visual and auditory saliency maps is performed,

which results in an increase in the weight of the auditory saliency map and a reduction in weights for the individual feature conspicuity maps of the visual domain.

In the last method, in addition to a linear combination of the visual and auditory saliency maps, a product term is added as,

$$\begin{aligned}\mathcal{AVSM}_3(x,y,t) = \Big(&\mathcal{R}(\mathcal{VSM}(x,y,t)) + \mathcal{R}(\mathcal{A}(x,y,t)) \\ &+\mathcal{R}(\mathcal{VSM}(x,y,t)) \otimes \mathcal{R}(\mathcal{A}(x,y,t))\Big)\end{aligned} \tag{7.10}$$

where the symbol, $\otimes$ denotes a point-wise multiplication of pixel values of the corresponding saliency maps. The effect of the product term is to increase the saliency of those events that are salient in both visual and auditory domains, thereby to enhance the saliency of spatiotemporally concurrent audiovisual events. A comparison of the different saliency maps in detecting salient events is detailed in Section 7.4.

7.3 Data and Methods

Audio-Visual data is collected using the VisiSonics RealSpace^{TM} audio-visual camera [253, 255]. The AV camera consists of a spherical microphone array with 64 microphones arranged on a sphere of 8 inches diameter and 15 High-Definition (HD) cameras arranged on the same sphere (Figure 7.4). Each video camera can record color (RGB) videos at a resolution of 1328×1044 pixels per frame and 10 frames per second for different directions on the sphere. The microphones record high fidelity audio at a sampling rate of 44.1 kHz per channel. The audio and video data are converted into a single USB 3.0 compliant stream

which is accepted by a laptop computer with Graphical Processing Units. The 15 videos are stitched together to produce a panoramic view of the scene in two different projection types: spherical and Mercator. The audio and video streams are synchronized by an internal Field Programmable Gate Array (FPGA) based processor. The equipment can be used to localize sounds and display them on the panoramic video in real time and also record AV data for later analysis. The length of each recording can be set at any value between 10 seconds and 390 seconds . The gain for each recording session can set to three predefined values: -20 dB, 0 dB and +20 dB. This is particularly helpful to record sounds with high fidelity in indoor, outdoor and noisy conditions.

To compute the loudness and location estimate of sound sources in the scene, Spherical Harmonics Beamforming (SHB) technique is used. The 3D sound field sampled at discrete locations on a solid sphere is decomposed into spherical harmonics, whose angular part resolves the direction of the sound field. Spherical harmonics are the 2D counterpart of Fourier transforms (defined on a unit circle) defined on the surface of a sphere. A description of the mathematical details of the SHB method is beyond the scope of this thesis, interested readers can find details in [253, 254, 255].

The 64 audio channels recorded at 44100 Hz are divided into frames, each of 4410 samples. This gives us 10 audio frames per second, which is equal to the video frame rate. Spherical harmonic beamforming is done for the audio frames to locate sounds in the frequency range, $[300, 6500]$ Hz. For natural sounds, a wide frequency range, as chosen is sufficient to estimate location and loudness of most types of sounds including speech and music. The azimuth angle for SHB is chosen in the range, $[0, 2\pi]$ with the angular resolution

Figure 7.4: The Audio-Visual Camera being used to collect data in a recording session outdoors

of $\frac{2\pi}{128}$ radians, *i.e.* , 2.8125^0 and the elevation angle in $[0, \pi]$ with an angular resolution of $\frac{\pi}{64}$ radians, *i.e.* , 2.8125^0 radians. The output of SHB is the auditory location and loudness estimate map as shown in Figure 7.3.

We collected four audiovisual datasets using the AV camera equipment, where three datasets are 60 seconds in length and the other one is of 120 seconds duration, all indoors. The AV camera equipment and our algorithms can handle data from any type of audiovisual surroundings, indoor or outdoor. We made sure all combination of salient events in purely visual intensity, color, motion, audio and audiovisual domains were present. SHB was performed with parameters set as explained in the previous paragraph to get the sound location and loudness estimates. The videos are stitched together to produce panoramic image in the Mercator projection which is used in our saliency computation. The stitch depth for panoramic images was 14 feet, hence in objects that are too close to the AV camera appear to be blurred due to overlapping of images from different cameras on the sphere (Figure 7.5).

The scene consisted of a loudspeaker placed on a desk in one corner of the room (green box) playing documentaries, a person (author) either sitting or moving around the AV camera equipment clapping or uttering numbers out loud, an air conditioning vent (blue box) making some audible noise and other objects visible in the scene. The loudspeaker acts like a stationary sound source, which was switched on/off during the recording session to produce abrupt onsets/offsets. The person wearing a black/blue jacket moving around the recording equipment acts like a salient visual object with or without motion, when speaking acts as an audiovisual source. The person moves in and out of the room producing

abrupt motion onsets/offsets. The air conditioning vent, which happens to be very close to the recording equipment is a source of noise which is audible to anyone present in the room, acts as another stationary audio source. The bright lights present in the room act as visually salient stationary objects. The set of sources together produce all possible combinations of visually salient events with or without motion, acoustically salient sources with or without motion and audiovisual salient events/objects with and without motion. The dataset can be viewed/listened to at the following url: `https://preview.tinyurl.com/ybg4fch4`. Results of audiovisual saliency, comparison with unisensory saliencies on this dataset are discussed in the next section.

7.4 Results and Discussion

First, we will examine which of the three audiovisual saliency computation methods described in Section 7.2.6, Eqs 7.8 - 7.10 performs well for different stimulus conditions. Then we will compare results from the best AVSM with the unisensory saliency maps followed by discussion of the results.

All saliency maps computed as explained in Section 7.2.6 will have salience value in $[0, 1]$ range. On such a saliency map, unisensory or audiovisual, anything above a threshold of 0.75 is determined as highly salient. This threshold is same for all saliency maps, Visual Saliency Map ($\mathcal{VSM}$, variables (x, y, t) dropped as unnecessary here), Auditory Saliency Map ($\mathcal{ASM}$) and the three different Audio-Visual Saliency Maps ($\mathcal{AVSM}_i$, where $i = 1, 2, 3$). Hence, this provides a common baseline to compare the performance of unisensory SMs with AVSM, and among different AVSMs.

Figure 7.5: The audiovisual data collection scene. The scene consists of a loudspeaker (green box), a person and an air conditioning vent (blue box)

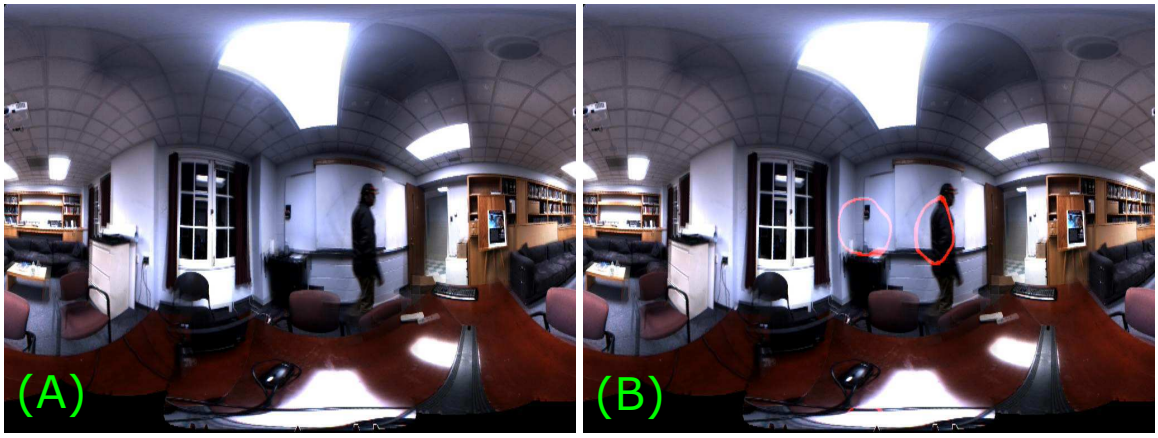

Figure 7.6: Visualization of results with isocontours. (A) Input frame # 77 of Dataset 2. (B) The red contour superimposed on the same input frame is the isocontour of salience values. All the values along the red line are equal to the threshold value. Anything inside the closed red isocontour has a salience value greater than 0.75

To visualize the results we did the following: On the saliency map (can be $\mathcal{VSM}$, $\mathcal{ASM}$ or $\mathcal{AVSM}_i$), saliency value based isocontours for the threshold of 0.75 are drawn and superimposed on each of the input video frame. For example, see Figure 7.6, where $\mathcal{AVSM}_1$ for frame # 77 of Dataset 2 is shown. Any thing that is inside the closed red contour of Figure 7.6(B) is highly salient and has a saliency value greater than 0.75. Outside the isocontour, the salience value is less than 0.75. Exactly, along the isocontour the salience value is 0.75 (precisely, 0.75 ± 0.02).

The results can be best interpreted by watching the input and different saliency map videos. But, since it is not possible to show all the frames and for the lack of a better way of presenting the results, we display the saliency maps for a few key frames only. The videos and individual frames of the saliency maps are available at the url: `https://preview.tinyurl.com/ybg4fch4`.

Figure 7.7 shows $\mathcal{AVSM}_1$, $\mathcal{AVSM}_2$ and $\mathcal{AVSM}_3$ for input image frame # 393 of

Dataset 1. At that moment in the scene, the loudspeaker (at the center of the image frame) was playing a documentary and the person was moving forward. So, there is a salient stationary auditory event and a salient visual motion. From visual inspection of Figure 7.7, it is clear that $\mathcal{AVSM}_1$, $\mathcal{AVSM}_2$ and $\mathcal{AVSM}_3$ give roughly the same results, and are able to detect salient events in both modalities. This is the behavior we see in all AVSMs ($\mathcal{AVSM}_i$) for a majority of frames. But, in some cases, when the scene reduces to a static image, the behavior exhibited by each of the methods will be somewhat different.

Consider, for example, frame # 173 of Dataset 3, where the visual scene is equivalent to a static image with a weak auditory stimulus, which is the air conditioning vent noise (Figure 7.8). Here, according to $\mathcal{AVSM}_1$ (Figure 7.8 (B)), the most salient location coincides with the strongest intensity based salient location at the bottom part of the image. This is because in $\mathcal{AVSM}_1$ we averaged the conspicuity maps with equal weights. So, when salient audio or motion is not present, the AVSM automatically switches to being a static saliency map with Color, Intensity and Orientation as dominant features. But, in $\mathcal{AVSM}_2$ it is computed as the average of visual and auditory saliency maps, hence it leads to redistribution of weights in such a way that each of the visual features contributes only one-eighth to the final saliency map and audio channel contributes one half. As a result, auditory reflections could get accentuated and show up as salient, which may not match with our judgment, as seen in Figure 7.8 (C). In reality, such auditory reflections are imperceptible, hence may not draw our attention as indicated by $\mathcal{AVSM}_2$.

$\mathcal{AVSM}_3$, in which a multiplicative term $\mathcal{VSM}(x, y, t) \times \mathcal{ASM}(x, y, t)$ is added, accentuates the conjunction of visual and auditory salient events if they are spatio-temporally

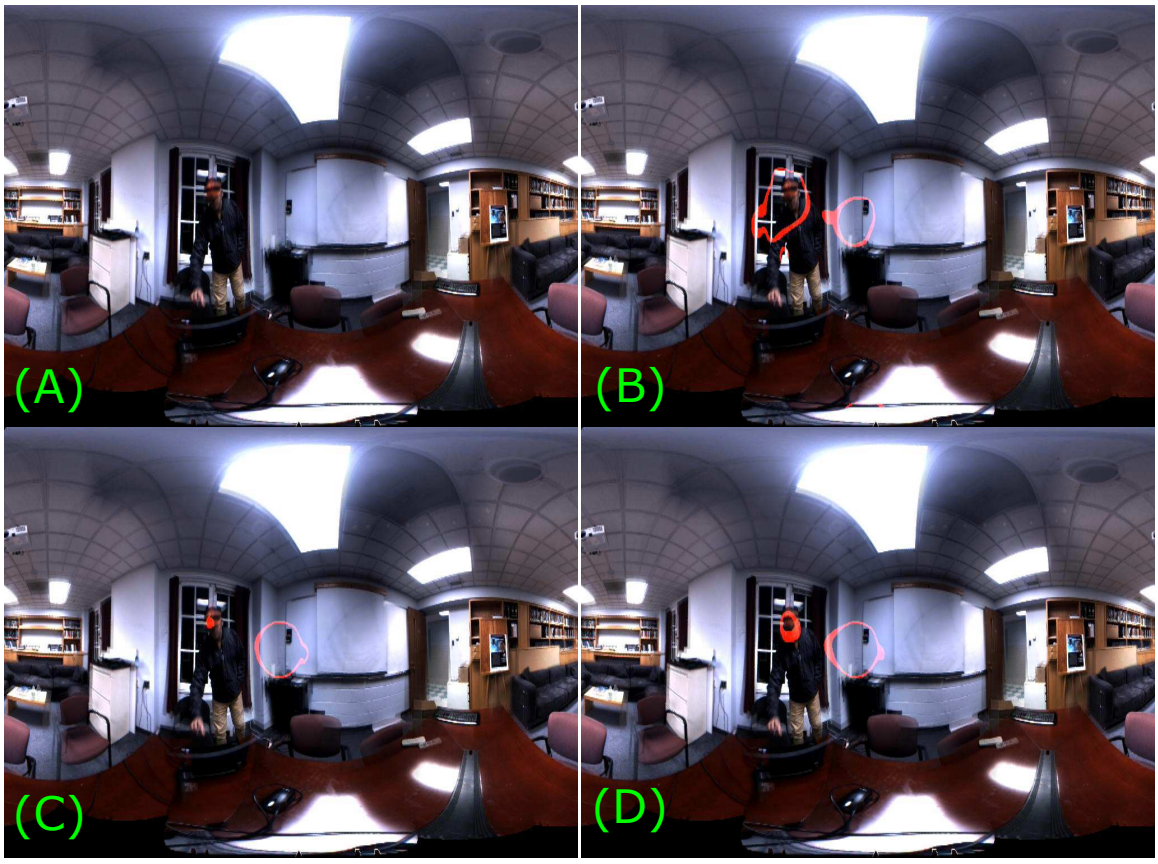

Figure 7.7: Comparison of audiovisual saliency computation methods. (A) Input frame # 393 of Dataset 1. A small loudspeaker at the corner of a room, which is located almost at the center of the image frame, is playing a documentary, hence constitutes a salient stationary auditory event. The person is leaning forward gives rise to salient visual motion. (B) $\mathcal{AVSM}_1$ computed using Eq 7.8 where all 5 feature channels are combined linearly with equal weights. (C) $\mathcal{AVSM}_2$ computed using Eq 7.9 where $\mathcal{VSM}$ and $\mathcal{ASM}$ are averaged. (D) $\mathcal{AVSM}_3$ computed using Eq 7.10. All methods give similar results with minor differences

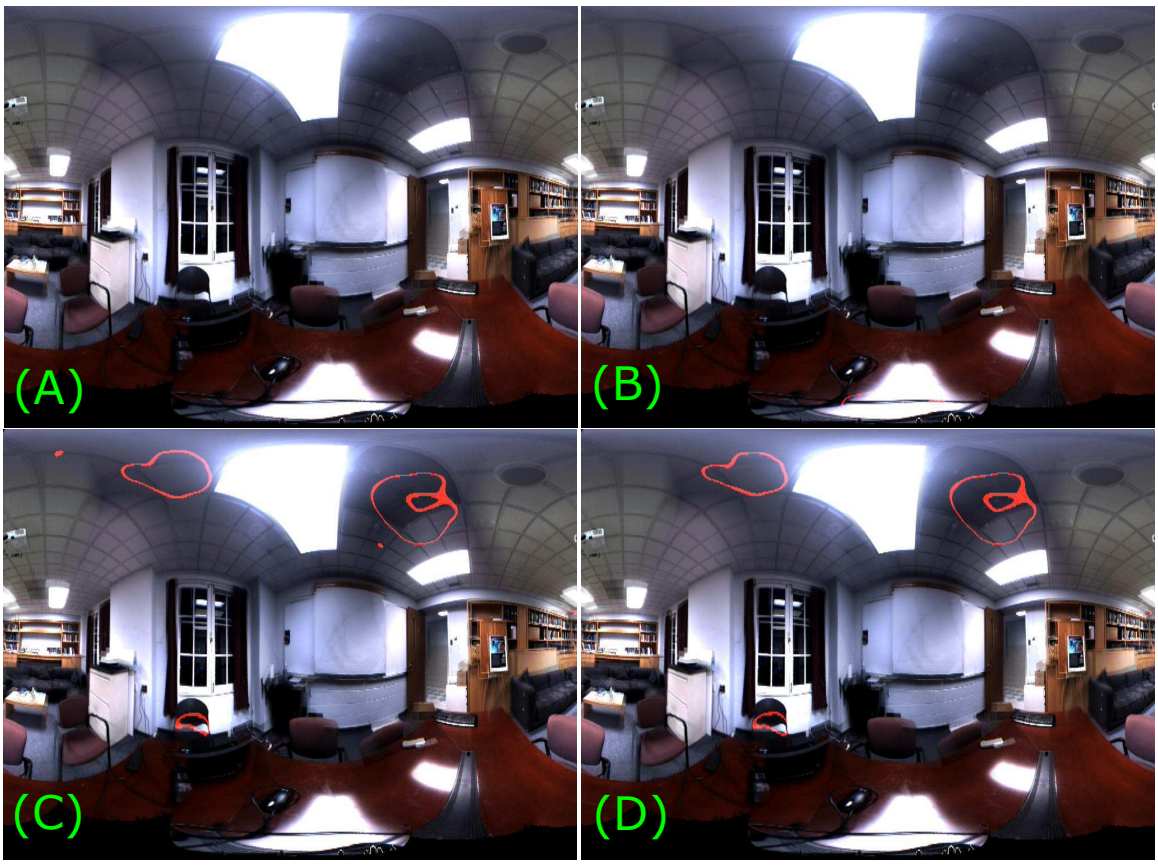

Figure 7.8: Audiovisual saliency in a static scene. (A) Input frame # 173 of Dataset 3. The scene is almost still reducing the visual input to a static image with a weak auditory noise emanating from the air conditioning vent. (B) The most salient location according to $\mathcal{AVSM}_1$ coincides with the intensity based salient location at the bottom of the image. (C) $\mathcal{AVSM}_2$ shows some locations that are not salient according to any feature. This may be happening due to exaggeration of auditory reflections which are detected as salient in $\mathcal{AVSM}_2$. (D) $\mathcal{AVSM}_3$ shows similar results as $\mathcal{AVSM}_2$

coincident. But, since auditory and visual saliencies already contribute equally instead of the five independent features making equal contributions, the effect of the multiplicative term is small, so we see that $\mathcal{AVSM}_3$ has similar behavior as $\mathcal{AVSM}_2$. We did not investigate whether the conjunction of individual feature conspicuity maps, like $\mathcal{I}(x,y,t) \times \mathcal{ASM}(x,y,t)$, $\mathcal{O}(x,y,t) \times \mathcal{C}(x,y,t)$, *etc* can result in a better saliency map. But based on visual comparison we can conclude that $\mathcal{AVSM}_1$, where each feature channel contributes equally, irrespective of whether it is visual or auditory, is a better AVSM computation method compared to $\mathcal{AVSM}_2$ and $\mathcal{AVSM}_3$.

So an important observation we can make at this point is that, even though vision and audition are two separate sensory modalities and we expect them to equally influence the bottom-up, stimulus driven attention, this may not be the case. Instead, we can conclude that each feature irrespective of the sensory modality makes the same contribution to the final saliency map from a bottom-up perspective.

Next, we will compare how $\mathcal{AVSM}_1$ performs in comparison to unisensory saliency maps, namely $\mathcal{VSM}$ and $\mathcal{ASM}$. Since $\mathcal{AVSM}_1$ was found to be better, the other two AVSMs are not discussed.

Figure 7.9 shows $\mathcal{ASM}$, $\mathcal{VSM}$ and $\mathcal{AVSM}_1$ for frame # 50 of Dataset 4, where the person moving is the most salient event, which is correctly detected in $\mathcal{VSM}$ and $\mathcal{AVSM}_1$, but not in $\mathcal{ASM}$. This is expected.

Next, in Figure 7.10 saliency maps for frame # 346 of Dataset 2 are shown, where audio from the loudspeaker is the most salient event, which is correctly detected as salient in $\mathcal{ASM}$ and $\mathcal{AVSM}_1$, but missed in $\mathcal{VSM}$, which agrees with our judgment.

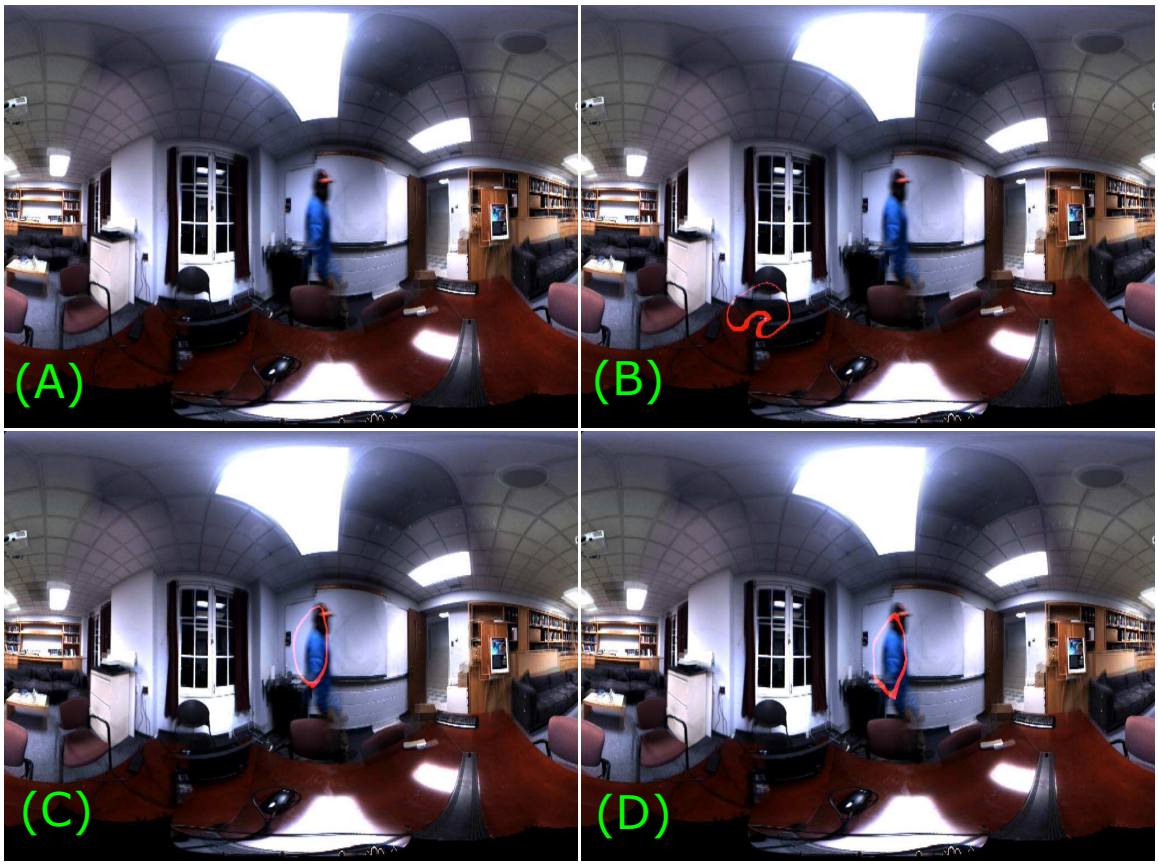

Figure 7.9: Comparison of AVSM with unisensory saliency maps (A) Input frame # 50 of Dataset 4. The most prominent event in the scene is the person moving. (B) The most salient location according to $\mathcal{ASM}$ misses the most salient location, but shows a different location as salient (C) $\mathcal{VSM}$ shows the prominent motion event as salient as expected (D) $\mathcal{AVSM}_1$ also captures the salient motion event as the most salient

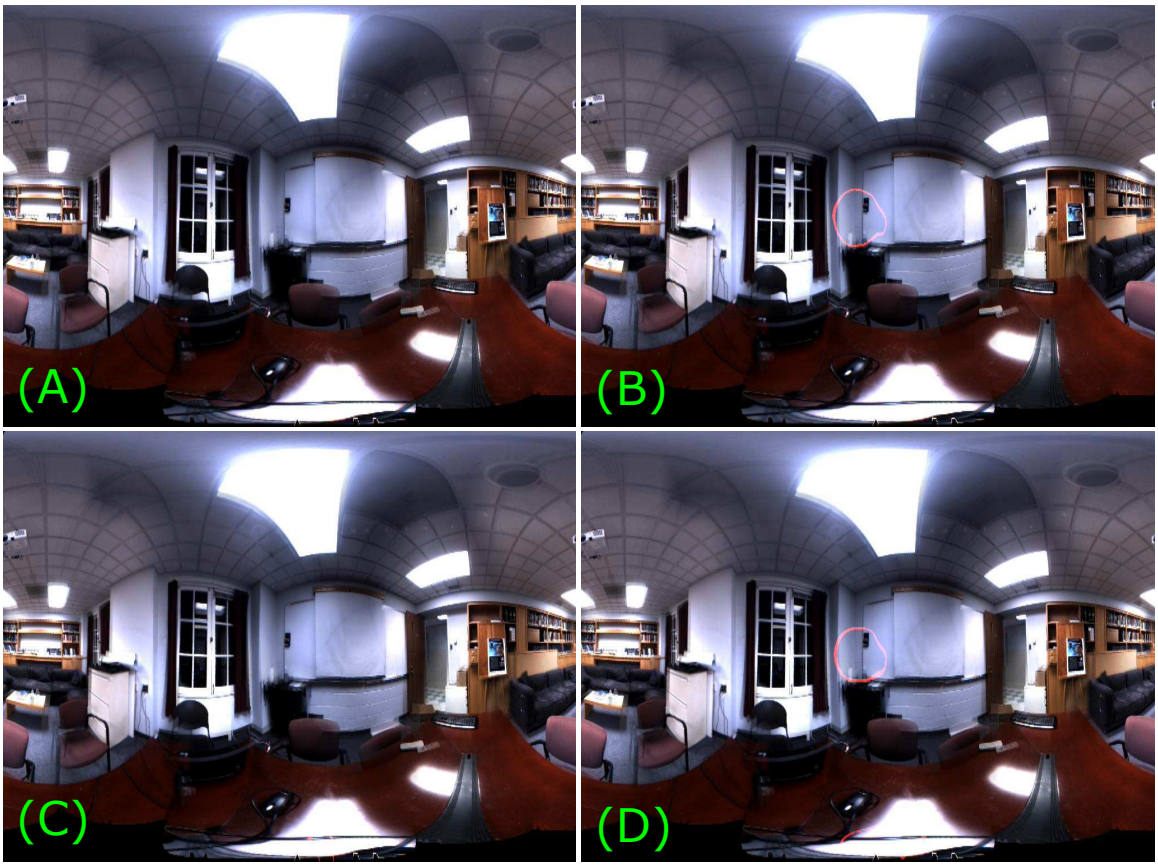

Figure 7.10: Comparison of AVSM with unisensory saliency maps (A) Input frame # 346 of Dataset 2. The most prominent event in the scene is the audio from the loudspeaker. (B) The audio event is salient in $\mathcal{ASM}$ (C) $\mathcal{VSM}$ in this case would be equivalent to a static saliency map, hence the auditory salient event is missed here(D) $\mathcal{AVSM}_1$ also captures the audio as the most salient as expected

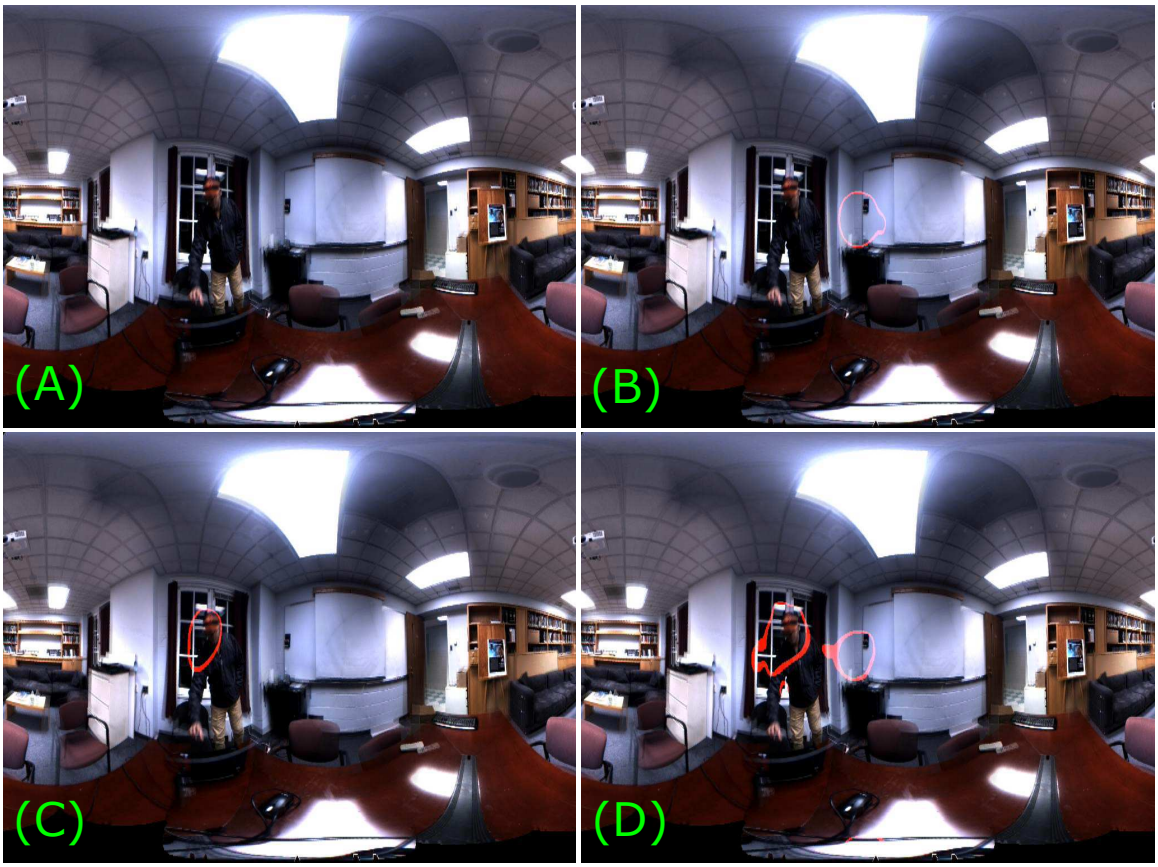

Figure 7.11: Comparison of AVSM with unisensory saliency maps (A) Input frame # 393 of Dataset 1. (B) $\mathcal{ASM}$ detects the salient auditory event, but misses the salient visual motion (C) $\mathcal{VSM}$ correctly detects the salient motion event, but misses the salient audio. (D) $\mathcal{AVSM}_1$ captures both valid salient events from two different sensory domains

In frame # 393 of Dataset 1, there is strong motion of the person as well as sound from the loudspeaker. The unisensory and audiovisual saliency maps are shown in Figure 7.11. Again, the salients events detected by the respective saliency maps agree with our judgment.

From these results we can conclude, the unisensory saliency maps detect valid unisensory events which agree with human judgment. At the same time, the audiovisual saliency map detects salient events from both sensory modalities, which again agree with our judgment. So, we can say, the AVSM detects more number of valid salient events compared to unisen-

sory saliency maps. The unisensory saliency maps miss the salient events from the other sensory modality. Hence, AVSMs in general, and $\mathcal{AVSM}_1$ in particular, perform better than unisensory saliency maps in detecting valid salient events. As a result, AVSMs can be more useful in a variety of applications like surveillance, robotic navigation, *etc.* Overall, the proto-object based audiovisual saliency map reliably detects valid salient events for all combinations of auditory, visual and/or audiovisual events in a majority of the frames. The readers can verify themselves additionally by watching the videos or looking at individual video frames at: `https://preview.tinyurl.com/ybg4fch4`.

An important distinguishing factor of our AVSM computation comes from the use of proto-objects. With proto-object based computation, we see that salient locations roughly coincide with object centers giving an estimate of audiovisual "objectness" [257]. So, this enables selection of image regions with possible objects based on saliency values. Moreover, since saliency gives a natural mechanism for ranking scene locations based on salience value, combined with "objectness" that comes from proto-objects, this can serve to select image regions for object recognition, activity recognition, *etc* with other methods, such as deep Convolutional Neural Networks [258, 259].

Second, due to linear combination of feature conspicuity maps, the model adapts itself to any scene type, static or dynamic scenes, with or without audio. Because of this, we get a robust estimate of bottom-up saliency in a majority of cases. Plus, the method works well for a variety of environments, indoor and outdoor.

Finally, the AVSM computed in this manner enables us to represent and compare saliencies of events from two different sensory modalities on a common scale. Other sensory

modalities or feature channels can be similarly incorporated into the model.

One of the factors that we have not considered in our model is the temporal modulation of audiovisual saliency. We treat each 100 ms interval as a snapshot, independent of previous frames and compute unisensory and audiovisual saliencies for each 100 ms frame. Even though this "memoryless" computation detects valid salient events well, temporal aspects are found to strongly influence saliency, especially from the auditory domain [27]. Hence, factoring in the temporal dependence of saliency can further improve the model. For example, in the few cases where saliency maps appears to be noisy, we can improve the results with temporal smoothing of the saliency maps. Though, the proportion of such noisy frames is very small compared to valid detections.

Temporal dependence of attention is important from the perspective of perception as well. For example, a continuous motion or an auditory alarm can be salient at the beginning of the event due to abrupt onset, but if it continues to persist, we may switch our attention to some other event, even though it is prominent in the scene. The mechanism and time course of multisensory attentional modulation needs to be further investigated and incorporated into the model.

Another aspect, related to temporal modulation of audiovisual saliency that we have not considered is the Inhibition of Return (IOR) [21, 260]. IOR refers to increased reaction time to attend to a previously cued spatial location compared to an uncued location. The exact nature of IOR in the case of audiovisual attention is an active topic of research [261, 262]. More recent experimental evidence [261] suggests that IOR is not observed in audiovisual attention conditions. If this is the case, not having audiovisual IOR may not be a significantly

limiting factor, but certainly worth investigating.

Lastly, a drawback of our work is that the results are not validated with human psychophysics experiments. Since, saliency models aim to predict human attention based on bottom-up features, validating the results with human psychophysics experiments is necessary. Such validation would strengthen the findings of our study even more. But from visual judgment of the results, the readers can verify that the model is capable of selecting valid, perceptually salient audiovisual events for further processing. Moreover, our goal is to build a useful computational tool for automated scene analysis and the results show that the model is capable of doing so.

7.5 Conclusion and Future Work

We have shown that a proto-object based audiovisual saliency map detects salient unisensory and multisensory events, which agree with human judgment. The AVSM detects a higher number of valid salient events compared to unisensory saliency maps demonstrating the superiority and usefulness of proto-object based multisensory saliency map. Among the different audiovisual saliency methods, we show that linear combination of individual feature channels with equal weights gives the best results. The AVSM computed this way performs better compared to others in detecting valid salient events for static as well as dynamic scenes, with or without salient auditory events in the scene. Also, it is less noisy and more robust compared to other combination methods where visual and auditory conspicuity maps, instead of individual feature channels, are equally weighed.

In future, incorporating the temporal modulation of saliency would be considered. We

would also like to validate the AVSM with psychophysics experiments. Also, the role of Inhibition of Return in the case of audiovisual saliency map would also be investigated. In conclusion, a proto-object based audiovisual saliency map with linear and equally weighted feature channels detects a higher number of valid unisensory and multisensory events that agree with human judgment.

Chapter 8

Future Work

In the first part of the thesis (Chapters 1 - 5), with the objective of building a FGO having with local and global cues, we proceeded to identify new local cues of FGO. In Chapter 2, we established SA as a valid cue of FGO, which is robust to variations in patch size, image size, *etc.* We then showed with SVM based non-linear classifier trained on Spectral Anisotropy features, the FGCA can be improved from $\approx 60\%$ to $\approx 70\%$ in Chapter 3. In future, efforts should be focused on using more advanced machine learning/deep learning based training methods to improve FGCA even further. If the objective is to achieve best FGCA with SA features, deep learning, which is a more efficient method should be the natural choice.

In Chapter 4, biologically plausible SA computation was shown to be nearly as efficient as FFT based SA computation method of Chapter 2. We also showed biologically plausible SA computation is robust to variations in the number of orientations, scales, aspect ratio values, *etc.* However, we always used the Complex cell responses in all our computation.

It would be interesting to see if similar or better FGCA can be achieved with Simple Even or Odd cells alone. If this can be done, then the computational cost of computing SA would reduce by more than half. This would make the overall FGO model computation of Chapter 5 even more efficient, when SA computed from Simple Even or Odd cells can be incorporated into the model. Also, in Chapter 4, only two values of γ_{SA} were experimented with. Filter size increment was in steps of 2 pixels. A more careful tuning of all these parameters with fewer or higher number of scales, different filter size resolutions (for example, $9 \times 9, 10 \times 10, \ldots$ instead of $9 \times 9, 11 \times 11, \ldots$), aspect ratios, *etc* should be considered to either decrease the number of scales to reduce the computational cost or improve the FGCA even more. Careful parameter tuning to reduce the number of scales along with verifying whether SA can be computed with Simple Even or Odd cells alone without sacrificing FGCA can dramatically reduce the computational cost, an advantage in the FGO model of Chapter 5.

In Chapter 5, we added two local cues to the FGO model, SA and T-Junctions. For better performance, as discussed in previous paragraph, SA can be computed by finding best parameters for biologically plausible computation or training based methods (Chapter 3), and incorporated into the FGO model. Similarly, T-Junction determination based on SVMs or other training based methods should be explored.

As studying the effect of local cues, whether they can be useful was our focus in Chapter 5, the model parameters were not tuned for best FGCA. But the FGCA of the model can be improved by tuning the inhibitory weight, w_{opp} for each feature and each local cue and tuning feature specific weights in Eq 5.24. In addition, increasing the number of scales

in the model, having von Mises kernels, $\mathcal{CS}$ cells and $\mathcal{B}$ cells of multiple radii can all lead to even better FGCA. Having von Mises kernels, $\mathcal{CS}$ cells and $\mathcal{B}$ cells of multiple radii of multiple radii can help capture the convexity and surroundedness cues better. Also, the model's figure-ground response is computed by modulating the activity of $\mathcal{C}_\theta$ cells, which are computed using Gabor filter kernels. The response of $\mathcal{C}_\theta$ cells may not always exactly coincide with human drawn boundaries in the ground-truth, with which we compare the model's response to calculate FGCA. Hence, averaging the BO response in a small 2×2 pixel neighborhood and then comparing that with the ground-truth FG labels could yield improved FGCA. In future, we would like to explore these ideas in order to improve FGCA. Moreover, color based cues [263, 264], global cues such as symmetry [265] and medial axis [157] can be incorporated to improve the FGCA and make the model more robust.

From a computational cost perspective, image segmentation using the gPb [185] algorithm is the most expensive step in the FGO model with local cues. So, in order to decrease the computational cost, more efficient image segmentation algorithms should be explored. One efficient algorithm with similar performance as gPb (F-score, Arbelaez et al. [185] = 0.70 *vs.* F-score, Leordeanu et al. [186] = 0.69 on BSDS 500 dataset) by Leordeanu et al. [186] is a good candidate. Replacing gPb [185] algorithm with the algorithm by Leordeanu et al. [186] for image segmentation, hence T-Junction computation, can substantially reduce the computational overload, while achieving similar performance. Other recent methods with better image segmentation performance can also be considered if the cost is not too high.

As we explained in Chapter 5, Section 5.3 we restricted the model only upto the computation of BO responses, as this was the end point of FGO and the BO responses thus

obtained were compared against the BSDS figure/ground ground truth dataset to determine FGCA. But, we also mentioned that the model, after the addition of local cues, can still be used for the computation of proto-objects with Grouping cells ($\mathcal{G}$ cells in Figure 5.1), hence visual saliency as in Russell et al. [25]. With the addition of local cues, we see an improvement of 8.77% in the model in terms of FGCA. It would be interesting to see how this improvement in FGO translates into improvement in visual saliency. So, we strongly suggest carrying out the proto-object saliency computation as in Russell et al. [25], but with the addition of local cues. We should evaluate how the improved proto-object saliency model with local cues compares with the model of Russell et al. [25] in predicting eye fixations on standard saliency evaluation datasets.

Coming to the Audio-Visual Integration part (Chapters 6, 7), one factor that we did not consider while computing AVSM in Chapter 7 was the temporal dependence of saliency. In Chapter 7, we treated each 100 ms of audiovisual data as a snapshot, that did not have dependence on previous history. Even though this "memoryless" computation detects valid salient events well, temporal aspects are found to strongly influence saliency, especially from the auditory domain [27]. Hence, factoring in the temporal dependence of saliency can further improve the model. Temporal dependence of attention in audiovisual environments, important from the perspective of perception, should be further investigated and incorporated into the model.

The mechanism of Inhibition of Return in audiovisual environments, still unclear with contradicting results from psychophysics experiments (See Section 7.4 in Chapter 7 for a related discussion), should be studied and incorporated into the model.

Also, the AVSM computed in Chapter 7 was based on Russell et al. [25], devoid of local cues. We added visual motion and auditory location and loudness maps as two new features to the model of Russell et al. [25] to compute the audio-visual saliency map. It would be interesting to replace the Color, Orientation and Intensity channels in its current form of the model with those of Chapter 5, where local cues of Spectral Anisotropy and T-Junctions were added. Then compute AVSM with local cues to compare the differences. Also, the motion feature was computed as the magnitude of horizontal and vertical velocity estimates. Instead, we can use the more biofidelitic motion estimation method developed by Molin et al. [158]. In addition, motion based local cues of FGO, such as common fate motion [266], preference for advancing *vs.* receding motion [267] should also be incorporated.

Lastly, the results of Chapter 7 are not validated with human psychophysics experiments. Since, saliency models aim to predict human attention based on bottom-up features, validating the results by presenting it human beings and measuring how they allocate attention in such environments, how much of their attentional behavior in audiovisual environments can be explained purely based on bottom-up features should be studied to validate and further strengthen the findings of our study.

Appendix A

Chapter 2: Supplementary Information

A.1 Patch Extraction Procedure

As described in Section 2.3, for our analysis we require all image patches to have a common reference frame in which x varies along the OB and y varies orthogonal to it. Patches are therefore rotated into this common reference frame.

Let us denote the location on the OB between figure and ground (yellow dot in Fig. 2.1A) from which figure and ground patches are to be extracted, as $\mathbf{u}_{center}$. We select a point $\mathbf{u}_{tangent}$ on the tangent to the OB, with $\mathbf{u}_{tangent} \neq \mathbf{u}_{center}$. The distinction between figure and ground is made by visual inspection and a point $\mathbf{u}_{figure}$ is located anywhere on the figure side.

The angle θ_{rot} by which patch $\psi(x, y)$ is rotated is computed as,

$$\theta_{rot} = \angle \mathbf{v}_{FG} + \pi/2 \tag{A.1}$$

where the symbol $\angle \mathbf{v}$ stands for the angle between vector $\mathbf{v}$ and a fixed coordinate axis, say a horizontal border of the image. The vector $\mathbf{v}_{FG}$ is defined as,

$$\mathbf{v}_{FG} = \mathbf{u}_{tangent} + \lambda^*(\mathbf{u}_{center} - \mathbf{u}_{tangent}) \tag{A.2}$$

with λ^* is chosen to ensure that the projection of $\mathbf{u}_{figure}$ is located on the line segment connecting $\mathbf{u}_{center}$ and $\mathbf{u}_{tangent}$. Defining $\langle \mathbf{a}, \mathbf{b} \rangle$ as the scalar product of vectors $\mathbf{a}$ and $\mathbf{b}$, it is obtained from the quantity

$$\lambda = \frac{\langle (\mathbf{u}_{figure} - \mathbf{u}_{tangent}), (\mathbf{u}_{center} - \mathbf{u}_{tangent}) \rangle}{\langle (\mathbf{u}_{center} - \mathbf{u}_{tangent}), (\mathbf{u}_{center} - \mathbf{u}_{tangent}) \rangle} \tag{A.3}$$

as

$$\lambda^* = \max(0, min(\lambda, 1)) \tag{A.4}$$

A.2 Plots related to the LabelMe database

Section 2.5 in the main text showed results for the BSDS300 database, Figures 2.2 and 2.3. The corresponding plots for the LabelMe database are shown in Figures A.1 and A.2.

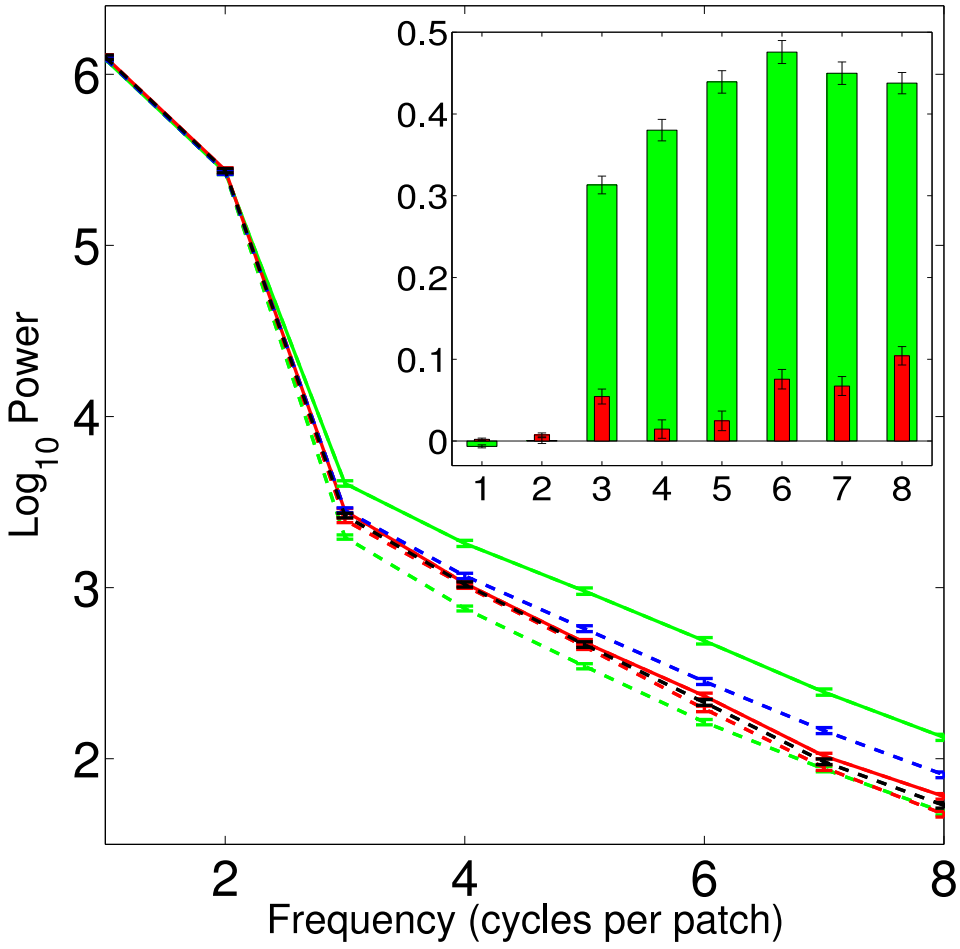

Figure A.1: Average power spectra of all patches of LabelMe data as function of spatial frequency. The unoriented spectra are represented by dashed blue (figure) and black (ground) lines. The oriented spectra in the plot are: $\overline{E}_{f\perp}$ (solid green line), $\overline{E}_{f\parallel}$ (dashed green line), $\overline{E}_{g\perp}$ (solid red line) and $\overline{E}_{g\parallel}$ (dashed red line). Inset: The difference in power ($\log_{10}(\bar{E}_{s\perp} - \bar{E}_{s\parallel})$) in each frequency bin. Axes same as in main figure. Green and red bars represent figure ($s = f$) and ground ($s = g$) differences respectively. Error bars are standard errors in figure and inset. For complementary BSDS results, see Figure 2.2

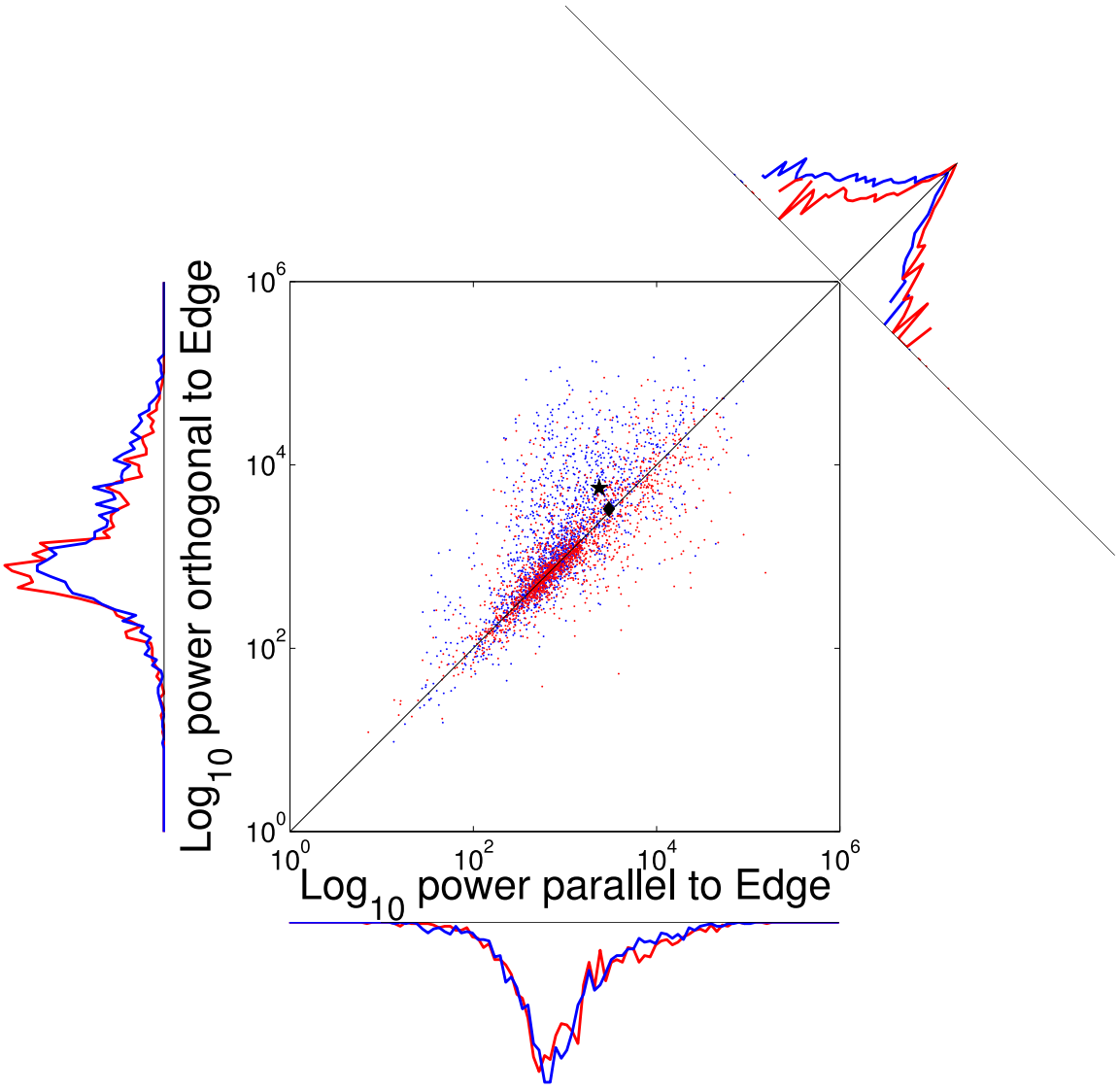

Figure A.2: Two-dimensional distribution of spectral power ($\log_{10} - \log_{10}$ axes) in bins 3–8 orthogonal *vs.* parallel to the OB for all LabelMe patches. Red, background side; ($[T_{g\perp}]_3^8$ *vs.* $[T_{g\parallel}]_3^8$), blue, figure side ($[T_{f\perp}]_3^8$ *vs.* $[T_{f\parallel}]_3^8$). The black diamond, very close to the identity line, shows the mean of the background. The black asterisk, above the identity, shows the mean of the figure. The distance between the figure-side mean and the identity line is larger compared to BSDS (Figure 2.3). The marginal distributions share their abscissas with the axes of the scatter plot and they have linear ordinates. The marginal distribution at the top right collapses data along the diagonal and has a logarithmic ordinate since the values of the central bins vastly surpass those of other bins. This marginal clearly shows the presence of spectral anisotropy, for BSDS, see Figure 2.3.

A.3 Maximum likelihood classification

Let the figure and ground parts of a patch pair be denoted by s_1 and s_2, where either s_1 or s_2 can be figure or ground. Let ρ_{s_1} and ρ_{s_2} be the SAs of s_1 and s_2 respectively. A patch pair can be in one of two configurations: (1) Correct (C=1), when figure is on top and ground is on bottom, and (2) Incorrect (C=0) when positions of figure and ground are reversed. Let s_1 and s_2 be the top and bottom parts, respectively, of a patch pair. Let γ be defined as the ratio of SAs of top and bottom parts, *i.e.* $\gamma = \frac{\rho_{s_1}}{\rho_{s_2}}$. The conditional distributions of γ for the two configurations (C=1 and C=0), also the *likelihoods* of the two configurations respectively are, $p(\gamma \mid C = 1)$ and $p(\gamma \mid C = 0)$.

The posterior probabilities of the two configurations are,

$$P(C = 1 \mid \gamma) = \frac{p(\gamma \mid C = 1)P(C = 1)}{p(\gamma \mid C = 1)P(C = 1) + p(\gamma \mid C = 0)P(C = 0)} \tag{A.5}$$

$$P(C = 0 \mid \gamma) = \frac{p(\gamma \mid C = 0)P(C = 0)}{p(\gamma \mid C = 1)P(C = 1) + p(\gamma \mid C = 0)P(C = 0)} \tag{A.6}$$

The correct configuration (C=1) is chosen when,

$$\frac{p(\gamma \mid C = 1)P(C = 1)}{p(\gamma \mid C = 1)P(C = 1) + p(\gamma \mid C = 0)P(C = 0)} > \frac{p(\gamma \mid C = 0)P(C = 0)}{p(\gamma \mid C = 1)P(C = 1) + p(\gamma \mid C = 0)P(C = 0)} \tag{A.7}$$

or,

$$\frac{p(\gamma \mid C=1)}{p(\gamma \mid C=0)} > \frac{P(C=0)}{P(C=1)} \tag{A.8}$$

But, we do not have any *a priori* reason to assume one of the configurations is more likely (hence prior probabilities, $P(C=0)$ and $P(C=1)$ will be the same). In that case, the choice is only based on the likelihoods, hence the test becomes a *maximum likelihood* test. Based on the *maximum likelihood* test, the correct configuration, C=1 is chosen when:

$$p(\gamma \mid C=1) > p(\gamma \mid C=0) \tag{A.9}$$

For the correct and incorrect configurations, the likelihood distributions are: $p(\gamma \mid C=1) = \frac{\rho_{s_1}}{\rho_{s_2}}$ and $p(\gamma \mid C=0) = \frac{\rho_{s_2}}{\rho_{s_1}}$, which are the distributions of ratios of SAs for the two configurations. So, Eq. A.9 can be written as,

$$\frac{\rho_{s_1}}{\rho_{s_2}} > \frac{\rho_{s_2}}{\rho_{s_1}} \tag{A.10}$$

which is equivalent to,

$$\rho_{s_1} > \rho_{s_2} \tag{A.11}$$

Hence, based on the *maximum likelihood* test, top part of the patch pair (s_1), is correctly chosen as figure and bottom as ground when $\rho_{s_1} > \rho_{s_2}$, which is the rule we use to perform figure-ground classification in Eq. 2.6 of Section 2.5.3.

patch size	Classification accuracy
18×18	60.63%
17×17	60.61%
16×16	62.57%
15×15	60.47%
14×14	61.96%

Table A.1: Classification accuracy *vs.* size of figure and ground patches.

A.4 Patch size *vs.* classification accuracy

All analyses in the main text were performed on figure and ground patches of size 16×16 pixels. The size of the patch in relation to the size of the object at the boundary from which it has been extracted determines the amount of local features contained in the patch. The BSDS300 database which has all images of same size, 481×321 is used in this section. In this analysis, we study the variation of figure/ground classification accuracy based on eq 2.6 with different sizes of figure and ground patches and hence the amount of local features on them.

We extract the figure and ground patches of 5 different sizes from 14×14 pixels up to 18×18 pixels in increments of one pixel. To eliminate confounding effects of size, once we extract a patch of a size different than 16×16 pixels, it is rescaled to 16×16 pixels, using linear interpolation of pixel values. Therefore, even though the patches are of the same size after re-scaling (16×16), they contain different densities of local features surrounding the boundary. Power spectra are computed for each dataset corresponding to 5 different sizes, as was done before in Sections 2.5.1 and 2.5.2. Our results indicate that the FG classification accuracy is little changed with ± 2 pixel change in size of the figure and ground patches (Table A.1).

Image size	Number of patch pairs	Classification accuracy
$0 - 0.5$	786	63.36%
$0.5 - 2$	360	68.14%
$2 - 4$	264	68.3%
> 4	348	67.05%

Table A.2: Classification accuracy *vs.* image size (in Mega Pixels) for the LabelMe dataset. Figure and ground patches are 16×16 pixels. The second column shows the numbers of figure/ground patch pairs available for the respective image sizes.

A.5 Image size *vs.* classification accuracy

In an approach that is complementary to that in A.4, we now measure the effect of image size on SA when patch size is kept constant, at 16×16 pixels. We use patches from the LabelMe dataset as it has a wide range of image sizes which we quantify in terms of the number of pixels present in the image.

From Table A.2, we see that image size only has a moderate impact on figure/ground classification accuracy. There is a tendency that image sizes in the range of 0.5 - 4 mega pixels give somewhat better classification results (68%) for patch size of 16×16 pixels than other image sizes. Our interpretation of these results is that when the images are very small (< 0.5 mega pixels), more global information is included in the SA analysis window, hence poorer classification accuracy results. On the other hand, when the image size is very large (> 4 mega pixels), classification accuracy is slightly reduced because the relatively small patch size compared to the foreground area makes it difficult to capture enough information about SA, since the variation in surface curvature of the underlying figure will be small.

A.6 SA of sharply focused patch pairs

In order to account for the photographer controlling the depth of field and excessively focusing on the foreground objects leaving the background smoothed and blurred out, we re-perform the entire analysis described in Sections 2.3 and 2.5 by discarding all blurry patches. Patch pairs were removed from the original datasets if either the foreground or background was blurred out due to limited depth of field. This yielded 1025 non-blurry patch pairs for BSDS and 1716 for LabelMe. The mean spectra corresponding to Figure 2.2 of the main text are plotted in Figures, A.3 and A.5 for LabelMe and BSDS respectively. Similarly, the plots corresponding to Figure 2.3 of the main text are shown in Figures, A.4 and A.6 respectively for LabelMe and BSDS. The statistical results for both databases are summarized below:

- Comparison of unoriented spectral power in bin 1 (figure *vs.* ground): Wilcoxon signed-rank tests (BSDS300: $p = 0.20$; LabelMe: $p = 0.66$)
- Comparison of $\bar{T}_f(1,8,1,8)$ *vs.* $\bar{T}_g(1,8,1,8)$: Wilcoxon signed-rank tests (BSDS300: $p = 0.30$; LabelMe: $p = 0.66$)
- Comparison of $\bar{T}_f(3,8,3,8)$ *vs.* $\bar{T}_g(3,8,3,8)$: Wilcoxon signed-rank tests (BSDS300: $p = 4.11 \times 10^{-14}$; LabelMe: $p = 4.63 \times 10^{-4}$)
- Comparison of $[T_{f\perp}]_3^8$ *vs.* $[T_{f\parallel}]_3^8$: Wilcoxon signed-rank tests (BSDS300: $p = 3.24 \times 10^{-23}$; LabelMe: $p = 1.49 \times 10^{-82}$)
- Comparison of $[T_{g\perp}]_3^8$ *vs.* $[T_{g\parallel}]_3^8$: Wilcoxon signed-rank tests (BSDS300: $p = 0.26$; LabelMe: $p = 0.69$)

- Figure-ground discrimination accuracy: (BSDS300: 63.41%; LabelMe: 64.16%)

A.7 Linear regression results

Please see Table A.3 for results related to linear regression of spectral powers when blurry patch pairs were not included in the analysis.

		slope(radians)	**CI(low)**	**CI (high)**	R^2
BSDS300	**Figure** (orthogonal *vs.*. parallel)	1.038	1.030	1.046	0.48
	Ground (orthogonal *vs.*. parallel)	1.00	0.994	1.007	0.66
LabelMe	**Figure** (orthogonal *vs.*. parallel)	1.071	1.064	1.079	0.53
	Ground (orthogonal *vs.*. parallel)	1.000	0.994	1.006	0.57

Table A.3: Regression of $\log_{10}$-transformed high-frequency spectral power in orthogonal and parallel orientations with slope as the only parameter for non-blurry patches only (analogous to Table 2.2 where all patches were included). Results for both datasets show the slope is higher in the foreground compared to background even after removing the blurry patch pairs. Again more oriented spectral power in orthogonal orientation in figure compared to the parallel orientation is observed. This indicates anisotropy of the figure cannot be caused by the photographer. Note that the confidence intervals of figure and ground are non-overlapping. Please see A.6 for other results when blurry patches were discarded from analysis.

A.8 Two dimensional spectra

An assumption underlying our work is that power varies in specific ways relative to the OB. To show that our observations are not due to some noise effect caused by random power fluctuations in arbitrary directions, we compute the 2D Fourier (power) spectra relative to the OB, separately for foreground and background patches. A 2D Hamming window is applied to each patch before DFT computation. The $\log_{10}$-transformed power spectra are

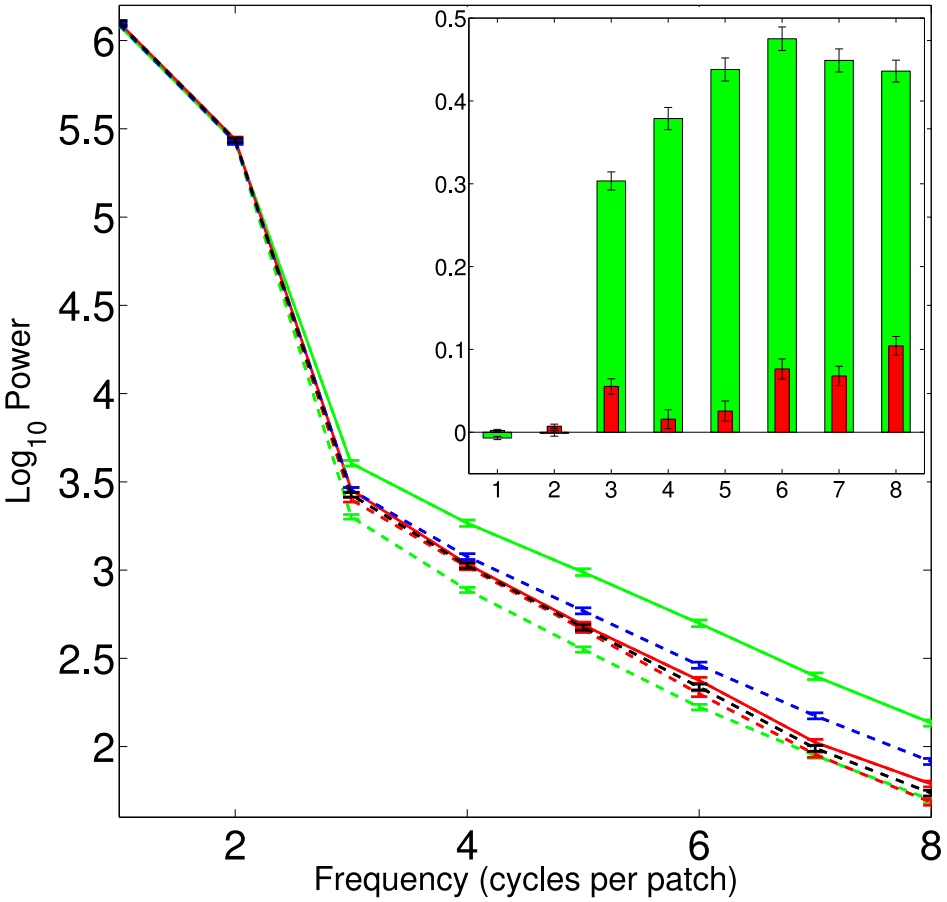

Figure A.3: Average power spectra of the 1716 non-blurry patch pairs of LabelMe dataset as function of spatial frequency. The unoriented spectra are represented by dashed blue (figure) and black (ground) lines. The oriented spectra in the plot are: $\overline{E}_{f\perp}$ (solid green line), $\overline{E}_{f\parallel}$ (dashed green line), $\overline{E}_{g\perp}$ (solid red line) and $\overline{E}_{g\parallel}$ (dashed red line). Inset: The difference in power ($\log_{10}(\bar{E}_{s\perp} - \bar{E}_{s\parallel})$) in each frequency bin. Axes same as in main figure. Green and red bars represent figure ($s = f$) and ground ($s = g$) differences respectively. Error bars are standard errors in figure and inset. Results from the BSDS database for non-blurry patches are similar, see Figure A.5.

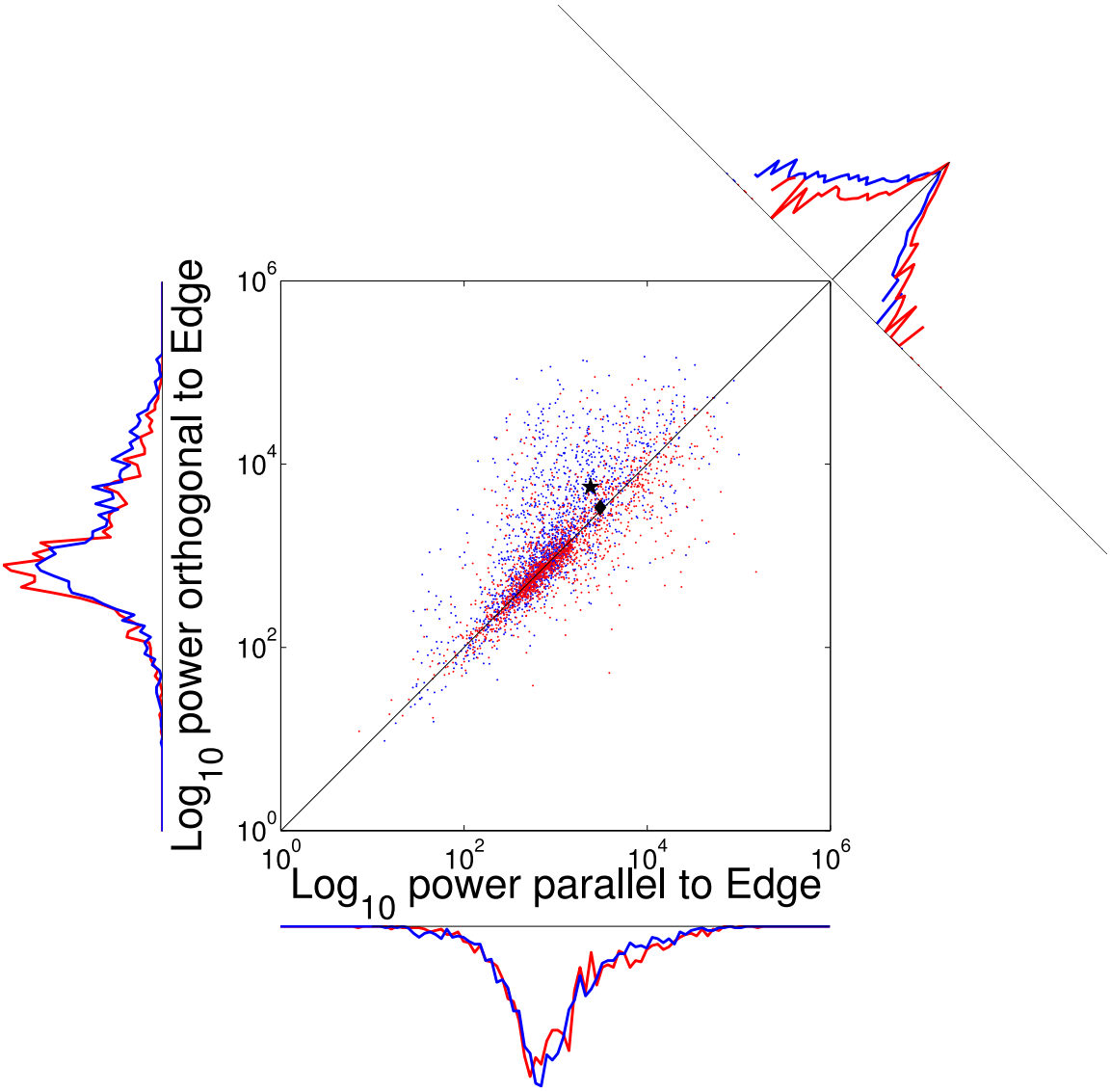

Figure A.4: Two-dimensional distribution of spectral power ($\log_{10} - \log_{10}$ axes) in bins 3–8 orthogonal *vs.* parallel to the OB for 1716 non-blurry LabelMe patches. Red, background side; ($[T_{g\perp}]_3^8$ *vs.* $[T_{g\parallel}]_3^8$), blue, figure side ($[T_{f\perp}]_3^8$ *vs.* $[T_{f\parallel}]_3^8$). The black diamond, very close to the identity line, shows the mean of the background. The black asterisk, above the identity, shows the mean of the figure. The distance between the figure-side mean and the identity line is larger compared to BSDS (Figure A.6). The marginal distributions share their abscissas with the axes of the scatter plot and they have linear ordinates. The marginal distribution at the top right collapses data along the diagonal and has a logarithmic ordinate since the values of the central bins vastly surpass those of other bins. This marginal clearly shows the presence of spectral anisotropy, for BSDS, see Figure A.6.

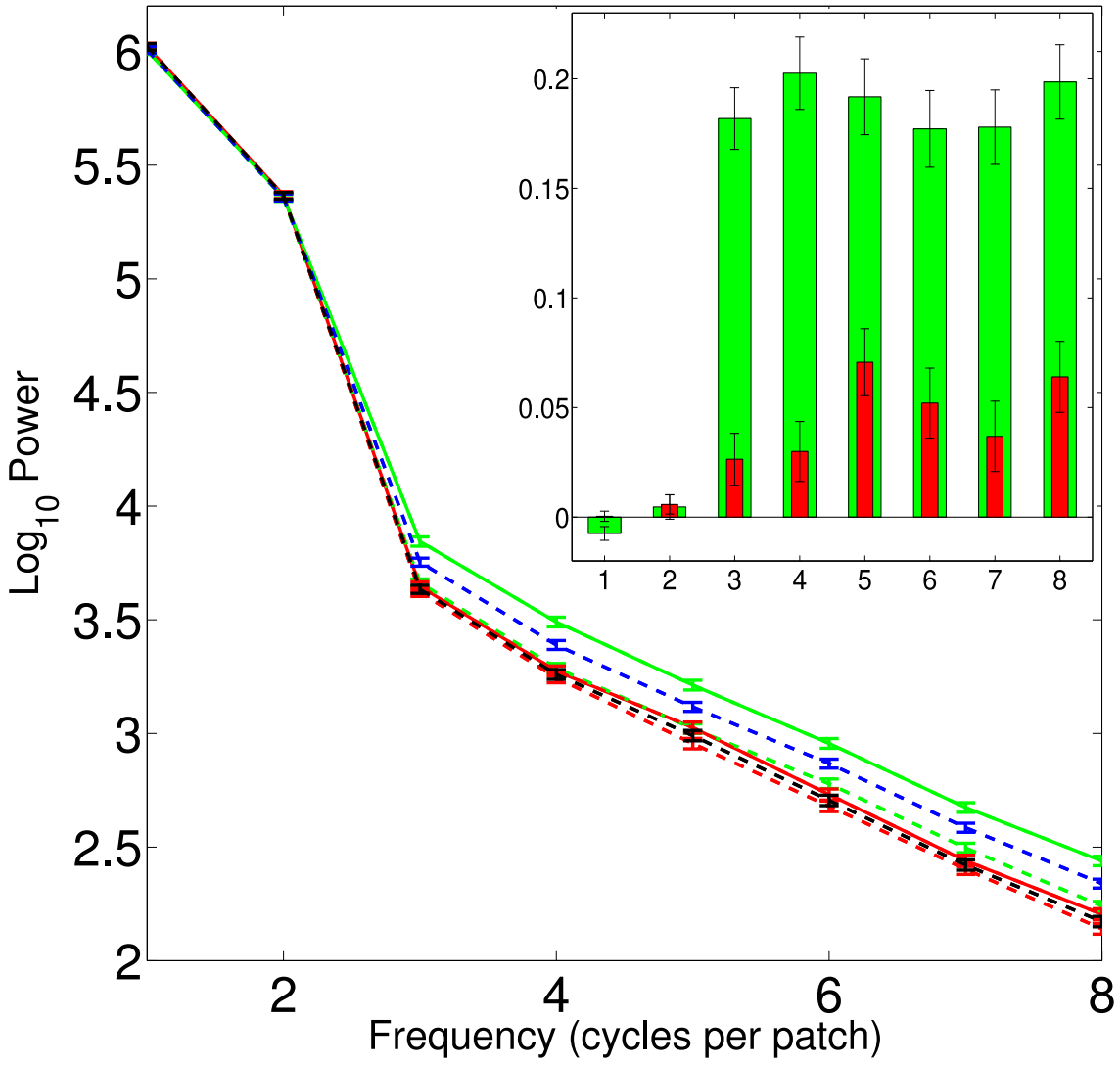

Figure A.5: Average power spectra of the 1025 non-blurry patch pairs of BSDS300 dataset as function of spatial frequency. The unoriented spectra are represented by dashed blue (figure) and black (ground) lines. The oriented spectra in the plot are: $\overline{E}_{f\perp}$ (solid green line), $\overline{E}_{f\parallel}$ (dashed green line), $\overline{E}_{g\perp}$ (solid red line) and $\overline{E}_{g\parallel}$ (dashed red line). Inset: The difference in power ($\log_{10}(\bar{E}_{s\perp} - \bar{E}_{s\parallel})$) in each frequency bin. Axes same as in main figure. Green and red bars represent figure ($s = f$) and ground ($s = g$) differences respectively. Error bars are standard errors in figure and inset. Results from the LabelMe database for non-blurry patches are similar, see Figure A.3.

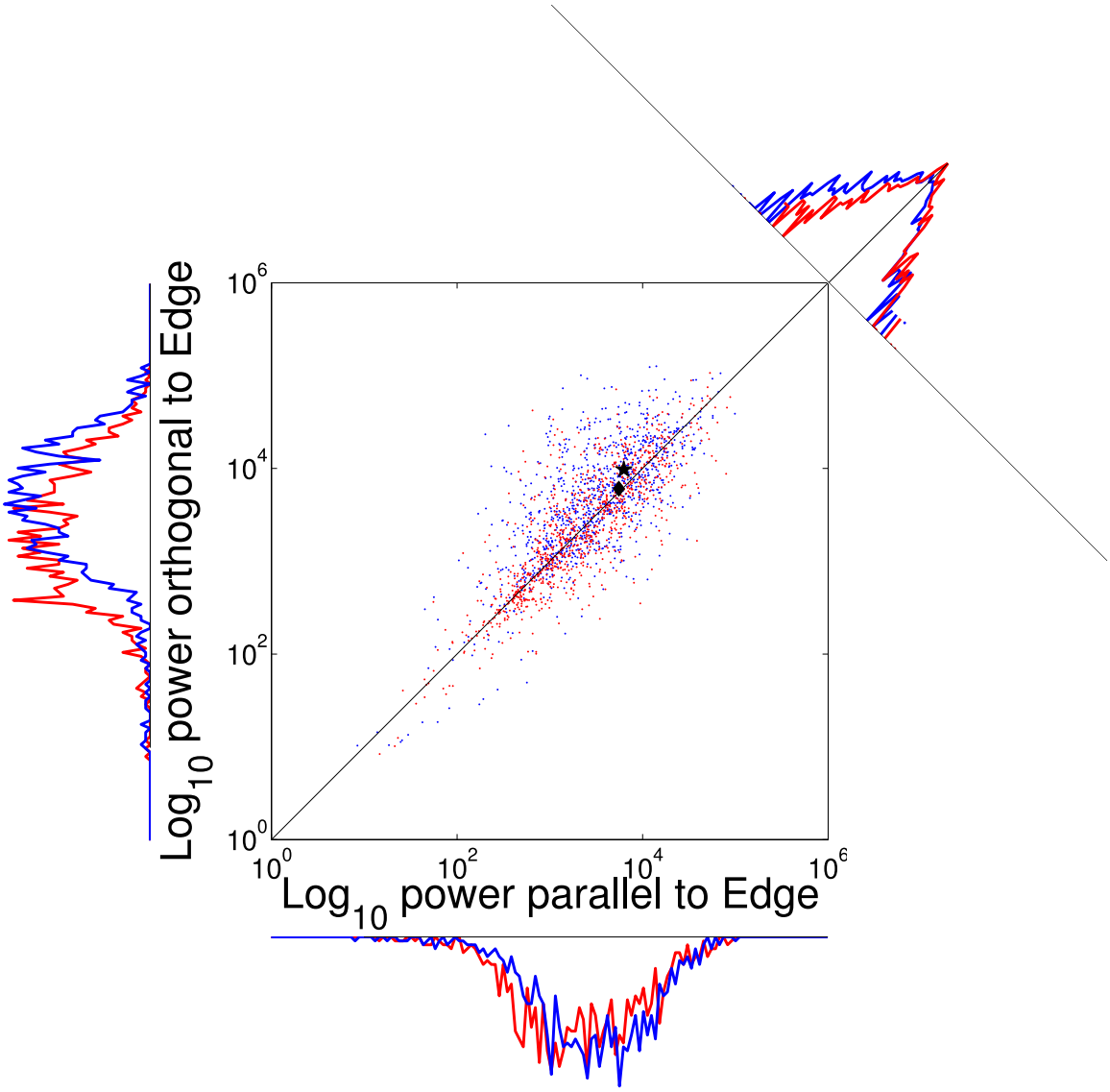

Figure A.6: Two-dimensional distribution of spectral power ($\log_{10} - \log_{10}$ axes) in bins 3–8 orthogonal *vs.* parallel to the OB for the 1025 non-blurry BSDS patches. Red, background side; ($[T_{g\perp}]_3^8$ *vs.* $[T_{g\|}]_3^8$), blue, figure side ($[T_{f\perp}]_3^8$ *vs.* $[T_{f\|}]_3^8$). The black diamond, very close to the identity line, shows the mean of the background. The black asterisk, above the identity, shows the mean of the figure. The distance between the figure-side mean and the identity line is even larger for LabelMe (Figure A.4). The marginal distributions share their abscissas with the axes of the scatter plot and they have linear ordinates. The marginal distribution at the top right collapses data along the diagonal and has a logarithmic ordinate since the values of the central bins vastly surpass those of other bins. This marginal clearly shows the presence of spectral anisotropy, and again the effect is stronger in the LabelMe data, see Figure A.4.

averaged over all samples in each database (LabelMe: 1761; BSDS: 1475) to obtain the mean 2D spectra which are plotted in Figure A.7. The colormaps are scaled to enhanced visual clarity. The maximal power difference is observed between orthogonal and parallel orientations relative to the OB, and that is the case on the figure side only.

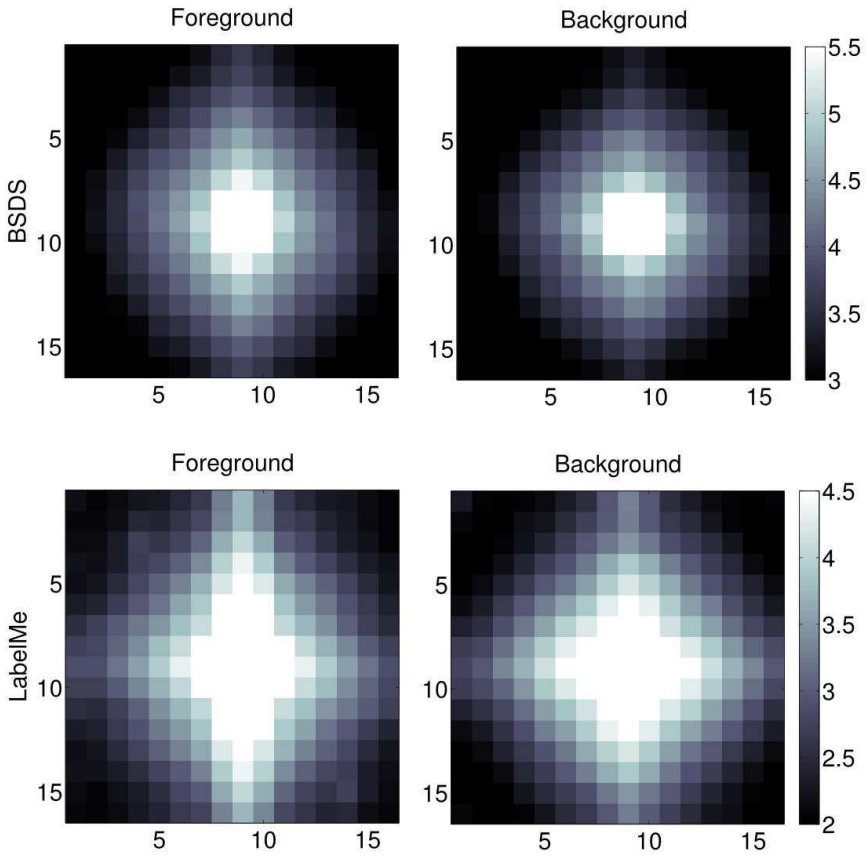

Figure A.7: Two dimensional power spectra ($\log_{10}$-transformed) of LabelMe (bottom two) and BSDS (top two) databases in figure (left) and ground (right). Same colormaps (scaled) are used for both figure and ground within each image set, but colormaps are different for different image sets. Patches are oriented as in Figure 2.1B. The spectral power is maximal along the vertical (corresponding to the orientation orthogonal to the occlusion boundary) in the figure. A minor anisotropy is noted even in the background which may be due to shadows cast by the foreground object, occasional minor errors in the labeling of occlusion borders, curvature effects of the boundary *etc.*

Appendix B

Chapter 5: Supplementary Information

B.1 Results with T-Junctions derived from ground-truth

When we added T-Junctions that were derived from the ground truth figure/ground labels, we saw a dramatic improvement in the FGCA of the model which is much higher than SA or what we obtained when T-Junctions were added based on automatically extracted edges or human drawn contours. This indicates, by taking the T-Junctions directly from the ground truth, we avoid the "inverted" T-Junctions, errors caused by mislabeled ground-truth and the effect T-Junctions on edge fragments that are not present in the ground truth figure/ground labeling. This is a case where the number of false positive T-Junctions is

FGCA of reference model	58.44%
FGCA of model with ground truth based T-Junctions	64.56%
Percentage Improvement	10.83%
Statistically Significant?	Yes
P-Value	0

Table B.1: Effect of T-Junctions when directly derived from figure/ground ground truth labelings: T-Junctions are added to the reference model. T-Junctions are derived from the figure/ground labels. The improvement in FGCA is higher than all other cases where a single local cue was added. This indicates if correctly incorporated T-Junctions are the strongest, unambiguous cue of FGO, but doing that based on a small local neighborhood is hard. Results for the test set of 100 images

zero and all the T-Junctions added are true positives. In such a case, even having a sparse number of T-Junctions (typically 3 - 10 per image) can lead to an improvement in FGCA that is dramatically higher than what we saw when SA alone was added. The FGCA achieved and results of right tailed, unpaired sample t-tests are tabulated in Table B.1. The optimal values of parameters were, $\alpha_{ref} = 0.0009$, $\alpha_{TJ} = 0.9991$ and $\alpha_{SA} = 0$. In the context of FGCAs we obtained when T-Junctions were derived by other methods, these results indicate: (i) T-Junctions can be strongest and most unambiguous cue of FGO, even though they are sparse if all of them match the ground truth, (ii) but reliably detecting the correct figure/ground relations at T-Junction locations based on a small local neighborhood algorithmically or for human observers can be hard for a variety of reasons that we have discussed before, which has been verified by psychophysics experiments too [190].

B.2 Some anomalies

First, we saw many instances of "inverted" T-Junctions, which are illustrated in Figure B.1. At an "inverted" T-Junction, the figure-ground relation indicated by the local

junction would be exactly contradicting the global configuration of objects, *i.e.* , the "stem" part of the T-Junction, which typically belongs to the background would be on the foreground side as per the global configuration of objects. But, when viewed locally, within a 15×15 pixels neighborhood, it appears to be a valid T-Junction. This problem has been discussed before by Palou and Salembier [171] as well.

In a few instances, the figure/ground relation indicated by the ground truth labels at the junction location do not match our judgment or what a "classical" T-Junction would indicate, illustrated in Figure B.2. At a T-Junction, the "stem" always belongs to the background side. But, at the T-Junctions illustrated in Figure B.2, this is not the case. Whereas the "inverted" T-Junctions arise due to the properties of the objects in the image, the discrepancies illustrated in Figure B.2 arise due to mislabeling of ground truth.

Lastly, the boundaries on which figure/ground labels are marked form only a subset of the human drawn contours or edges that would be typically detected by any segmentation algorithm. Some contours are selectively removed from the ground truth figure/ground label maps and there is no clear reasoning given by the database creators in [106], as to what type of human drawn contours were not labeled for figure/ground relations. A couple of examples are in Figure B.3, where you can see many edges that are in the human drawn contour map (last column), which would also be detected by any automated segmentation algorithm, are not present in the figure/ground ground truth map (middle column). Since T-Junctions are a by-product of edges, such T-Junctions derived either human drawn contours or automatically extracted edges, would impose a different figure-ground relationship, that need not necessarily agree with the ground truth. Hence, we could see reduced FGCA, even

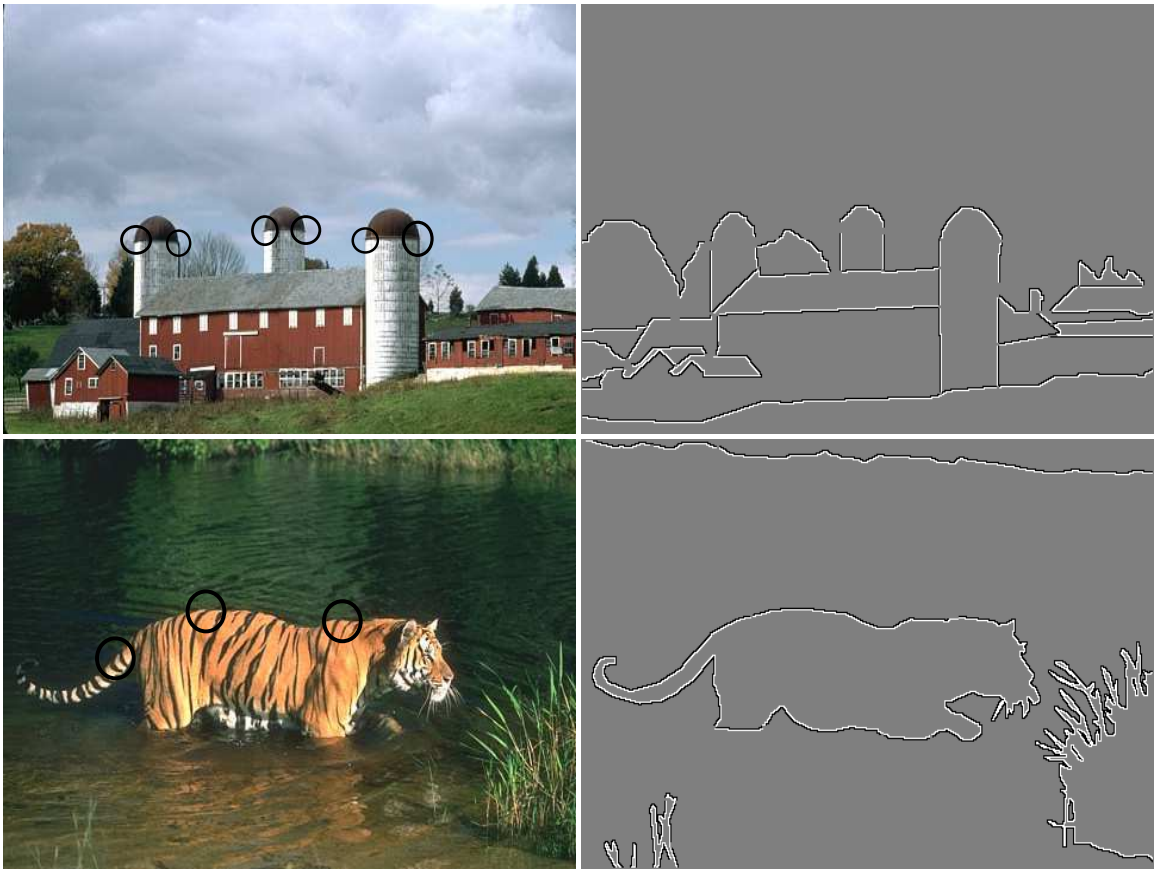

Figure B.1: Inverted T-Junctions, only a few are labeled with black circles surrounding them, but many instances exist in the images. Here, T-Junctions indicate a figure direction that locally makes sense, but points exactly opposite to the foreground(*i.e.* "stem" in the foreground) in the context of global configuration of objects. Left column: images with T-Junctions in black circles; Right column: ground truth figure/ground label map with white pixels on the figure side, black pixels on the background side

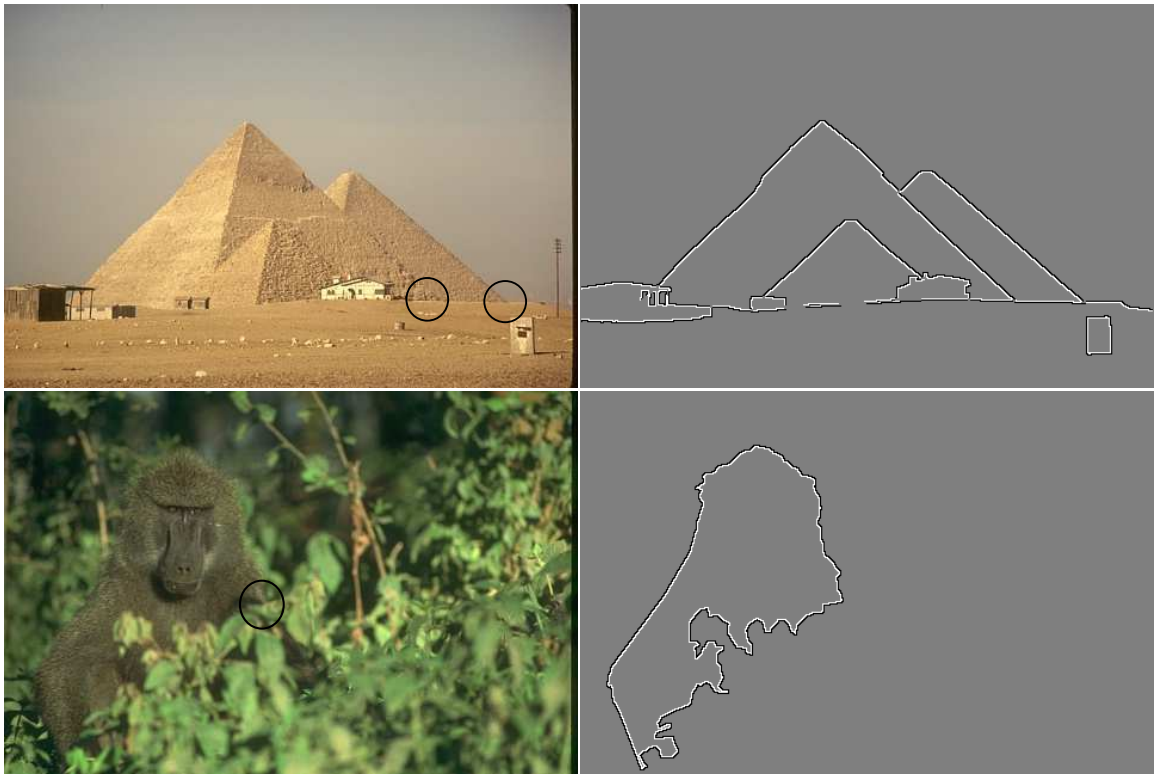

Figure B.2: The ground truth (Right column, with white pixels on the figure side, black pixels on the background side) appears to be mislabeled. In the monkey image, the shrub in front of the monkey is the foreground at T-Junction location (circled on the image). But, the ground truth indicates otherwise. Similarly, in the pyramids image, at the base of pyramids (see bottom-right part, circled), the ground surface in front of the pyramids should be foreground as it is nearer to the observer and occludes the base of the pyramid, but the ground truth label indicates otherwise

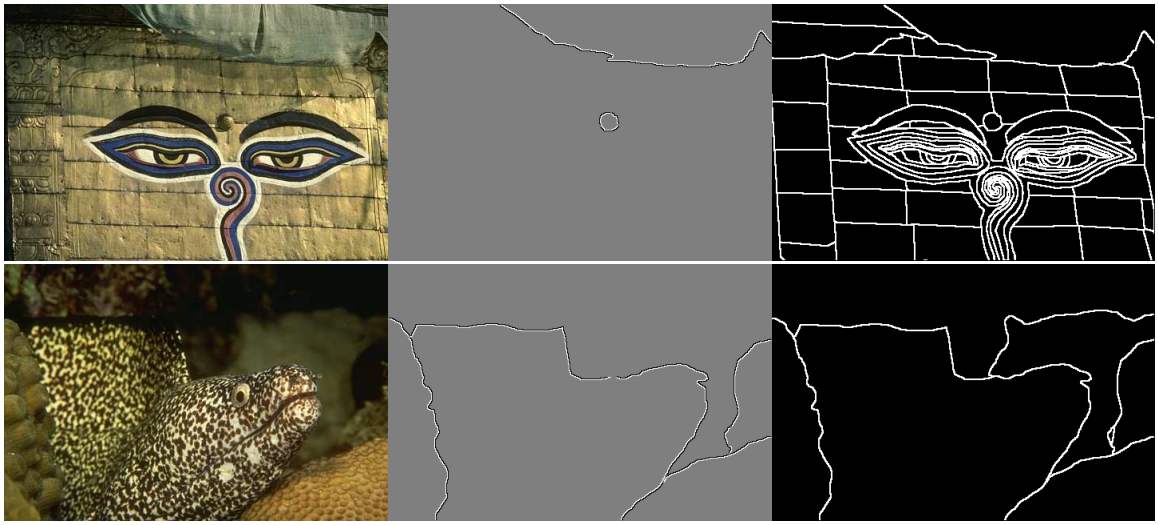

Figure B.3: Left column: Images; Right column: human drawn edges provided by BSDS dataset; Middle column: figure/ground ground truth (white pixels on the figure side, black pixels on the background side). Edges, even drawn by human labelers, are not present in the figure/ground ground truth. These edges would be detected by any image segmentation method, which produce T-Junctions whose influence may not agree with the ground truth

if the T-Junctions are correctly detected by an algorithm.

B.3 Computational Cost Analysis

We analyze the cost of computing the Figure Ground Organization model incorporating local and global cues. The objective is to measure the computatinal overhead of adding local cues into the FGO model that is devoid of any local cues, *i.e.* , the additional computational cost of adding local cues to the Reference model. The computational cost is measured in terms of the number of floating point operations (FLOPs) for an image being analyzed. The number of FLOPs for basic arithmetic operations (addition, subtraction, multiplication) are counted as 1 FLOP, comparison as 1 FLOP, division, square root of a number as 4 FLOPs, exponential, trigonometric functions as 8 FLOPs, *etc.* These estimates are based on [268].

The computational cost measured in FLOPs for an image for different versions of the model are summarized in this section. The image size is always, 321×481 pixels, the size of all images in the BSDS Figure/Ground dataset.

The analysis we present is a raw count of the FLOPs required to do certain computation. Actual implementation can have a reduced cost based on the algorithms used in MATLAB. For example, multiplying two $N_1 \times N_1$ matrices, takes N_1^3 multiplications and additions without any optimization (*schoolbook method*), but with optimized algorithms like Coppersmith-Winograd algorithm, it could be reduced [269] to $\mathcal{O}(N_1^{2.376})$. Since, a plethora of functions, optimized for the MATLAB platform are used in the actual computation, it is prohibitively time consuming to make an exact analysis of the actual computation cost. But, since all computation costs are measured based on the raw count of FLOPs, this allows comparison of the computational overhead involved in computing the local cues. Since our objective is to measure the computational overhead of adding local cues, we think it is justified to use this strategy in computing the computational cost.

Computation of SA involves filtering an image with Simple Even and Odd cells of different sizes, $9\times9, 11\times, 11, \ldots, 25\times25$, computing the Complex cell responses for 8 orientations at each scale by summing the Complex cell responses for the two BO directions. Then making a multiresolution pyramids for 10 scales.

Computing Simple cell response for each pixel involved F_s^2 adds and F_s^2 multiplications, where F_s is the Simple Even or Odd filter size. We need to do this for Simple Even and Odd cell types. So, that is $2 \times F_s^2 \times 2 = 4F_s^2$ FLOPs. Computing Complex cell response involves squaring each simple cell response (1 FLOP per cell type), summing them (1 FLOP) and

computing the square-root (4 FLOPs), total 7 FLOPs. This has to be done for 8 orientations and 2 BO directions. So, the total FLOP count will be $(4F_s^2 + 7) \times N_{ori} \times N_{BO}$ FLOPs, where $N_{ori} = 8$, is the number of orientations and $N_{BO} = 2$ is the number of BO directions per orientation. The total FLOP count for an image of size 321×481 pixels, the cost will be 28,251,677,376 FLOPs. This is the cost based on the current implementation where filters of different sizes are chosen. Instead, if we keep the filter size constant at 9×9 pixels, but do the same computation on multi-resolution image pyramid, the cost can dramatically reduce to less than 1,643,359,393 FLOPs. If we adopt the later strategy, it is important to remember SA is not scale invariant like other independent features in the model. Hence, SA should be computed only on the top 1 - 5 levels of the image pyramid. The effect of SA can still be reliably captured even if it is computed only at the top 1 - 5 layers, as it is robust to variation in image size (Section A.5).

Computing T-Junctions from an already computed edge map requires 49,250 FLOPs, assuming a maximum of 10 T-Junctions per image. Computing edge maps using the method of Arbelaez et al. [185] takes 21,367,708,791 FLOPs. This is based on the estimate in [270], that gPb algorithm requires about 138,391 FLOPs per pixel. The generalized boundary method of Leordeanu et al. [186] takes 16,151,464,747 FLOPs, based on our estimate.

The computation costs of different versions of the model, with/without local cues and the respective computational overhead are listed below:

- Reference model computation - no local cues: 133,787,537,703 FLOPs
- Reference model + SA (current implementation): 170,001,489,671 (computational overhead: 27.068%). Since we wanted to first study if SA is useful at all to begin

with, we attempted a method that we know achieves good results (based on Chapter 4 results). It would have become hard to study if we directly attempted the fixed filter size, multi-resolution image pyramid based SA computation. But, as we show in Chapter 4, even a single 9×9 filter based SA gives $\geq 59\%$ FGCA, so keeping the filter size fixed and computing SA based on multi-resolution image pyramid would yield essentially the same results as we see in Chapter 5. This would dramatically reduce the computational cost.

- Reference model + SA (ideal implementation): 143,393,171,688 FLOPs (computational overhead: 7.17%). In the ideal case, filter size would be kept constant and SA would be computed based on image pyramid. Moreover, by reducing the number of orientations to 4, instead of 8, the cost can be reduced by half to $\approx 3.5\%$. As we show in Chapter 4, biologically plausible SA computation is robust to variation in the number of orientations, it is possible to achieve similar improvement in FGCA ($\approx 7\%$) even with only 4 orientations. Additionally, only Simple cells can be used to reduce the computational cost even more (See Chapter 8).

- Reference model + T-Junctions (without edge segmentation step): 141,749,861,545 FLOPs (computational overhead: 5.95%)

- Reference model + T-Junctions (edges from Arbelaez et al. [185]): 163,117,570,336 FLOPs (computational overhead: 21.92%)

- Reference model + T-Junctions (edges from Leordeanu et al. [186]): 157,901,326,292 FLOPs (computational overhead: 18.02%).

The segmentation capabilities of Arbelaez et al. [185] and Leordeanu et al. [186] are very similar. On the same BSDS 500 dataset, Arbelaez et al. [185] achieves an F-score of 0.70, whereas Leordeanu et al. [186] achieves an F-score of 0.69.

- Reference model + SA + T-Junctions (actual implementation): 197,742,776,064 FLOPs (computational overhead: 47.8%)

- Reference model + SA + T-Junctions (ideal case): 165,918,214,037 FLOPs (computational overhead: 24.016%). In this case, edges required for T-Junction computation would be derived based on Leordeanu et al. [186] algorithm; SA would be computed with a fixed filter size of 9×9 pixels, but on multi-resolution image pyramid.

By computing SA at only 4 orientations, using only Simple Even or Odd cells instead of Complex cells, computing SA at only top 1 - 5 layers of the image pyramid and using an efficient image segmentation algorithm such as the one proposed by Leordeanu et al. [186], the ideal computational overhead of adding both SA and T-Junction local cues would be around 20% - 24%.

B.4 Local cues influencing only top 2 layers

We considered the question: Should the influence of local cues also be strictly local? Local cues, by definition, should be computed based on the analysis of a small patch of an image to determine figure-ground relations. This is what makes them computationally more efficient. But, should their influence also be local? There is no *a priori* reason why their influence should be strictly local. To verify whether there is higher benefit in adding

Model	**k = 2**	**k = 10**
Ref Model	-	58.44%
Ref + SA	62.42%	62.69%
Ref + T-Junctions (gPb edges)	59.12%	59.48%
Ref + T-Junctions (human labeled edges)	61.23%	61.98%

Table B.2: Local cues only at the top 2 layers: By adding each local cue only at the top 2 layers ($k = 2$), we see the FGCA we obtain is much lower than having them at all levels ($k = 10$).

them locally only at the top layer (*i.e.* , at native mage resolution only), we added them only at the top layer. For SA it resulted in a noticeable, but very small improvement. For T-Junctions, the change was barely noticeable. This could be due to extremely small size of von Mises filter kernels that we use ($R_0 = 2$ pixels) in comparison with the images size (481×321 pixels). So, we added the local cues to the top two layers. For each local cue added separately, the optimal parameters of the model were recomputed and those parameters were used to compute the FGCA. The versions of the model with local cues only at the top 2 layers did not give rise to better FGCA than what we saw earlier with the cues added at all scales. The results are summarized in Table B.2.

Bibliography

[1] M. Wertheimer. Untersuchungen zur Lehre von der Gestalt II. *Psychol. Forsch.*, 4: 301–350, 1923.

[2] Johan Wagemans. Perceptual organization. In *The Stevens' handbook of experimental psychology and cognitive neuroscience, sensation, perception and attention*, volume 2. John Wiley & Sons, Inc, Hoboken, NJ, 2017.

[3] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

[4] Johan Wagemans, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R Pomerantz, Peter A van der Helm, and Cees van Leeuwen. A century of gestalt psychology in visual perception: II. conceptual and theoretical foundations. *Psychological bulletin*, 138(6):1218, 2012.

[5] Celine Gillebert and GW Humphreys. Mutual interplay between perceptual orga-

nization and attention: a neuropsychological perspective. In *Oxford Handbook of Perceptual organization*, pages 736–757. Oxford University Press, 2015.

[6] F. T. Qiu, T. Sugihara, and R. von der Heydt. Figure-ground mechanisms provide structure for selective attention. *Nat. Neurosci.*, 10(11):1492–9, October 2007.

[7] E. Rubin. *Visuell wahrgenommene Figuren*. Kobenhaven: Glydenalske Boghandel, 1921.

[8] K. Koffka. *Principles of Gestalt psychology*. Harcourt-Brace, New York, 1935.

[9] P. Bahnsen. Eine Untersuchung uber Symmetrie und Asymmetrie bei visuellen Wahrnehmungen. *Zeitschrift fur Psychologie*, 108:129–154, 1928.

[10] S. E. Palmer. *Vision Science-Photons to Phenomenology*. MIT Press, Cambridge, MA, 1999.

[11] C.C. Fowlkes, D.R. Martin, and J. Malik. Local figure-ground cues are valid for natural images. *Journal of Vision*, 7(8), 2007.

[12] Friedrich Heitger, Lukas Rosenthaler, R von ver Heydt, Esther Peterhans, and Olaf Kübler. Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Research*, 32(5):963 – 981, 1992.

[13] P.S. Huggins, H.F. Chen, P.N. Belhumeur, and S.W. Zucker. Finding folds: On the appearance and identification of occlusion. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, 2:11718–11725, 2001.

[14] S.E. Palmer and T. Ghose. Extremal edges: A powerful cue to depth perception and figure-ground organization. *Psychological Science*, 19(1):77–84, 2008.

[15] S. Ramenahalli, S. Mihalas, and E. Niebur. Extremal edges: Evidence in natural images. In *IEEE CISS-2011 45th Annual Conference on Information Sciences and Systems*, pages 1–6, Baltimore, MD, 2011. IEEE Information Theory Society.

[16] L. Itti, G. Rees, and J. K. Tsotsos, editors. *Neurobiology of Attention*. Elsevier, San Diego, CA, 2005.

[17] Marisa Carrasco. Visual attention: The past 25 years. *Vision Research*, 51(13):1484 – 1525, 2011.

[18] J. K. Tsotsos and A. Rothenstein. Computational models of visual attention. *Scholarpedia*, 6(1):6201, 2011. revision #136393.

[19] C. Spence. Crossmodal attention. *Scholarpedia*, 5(5):6309, 2010. revision #75910.

[20] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiol.*, 4:219–227, 1985.

[21] L. Itti, C. Koch, and E. Niebur. A model of saliency-based fast visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

[22] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Neuroscience*, 2:194–203, 2001.

[23] R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1/2/3): 17–42, 2000.

[24] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, Nov 2006.

[25] Alexander F Russell, Stefan Mihalaş, Rudiger von der Heydt, Ernst Niebur, and Ralph Etienne-Cummings. A model of proto-object based saliency. *Vision research*, 94:1–15, 2014.

[26] S. Mihalas, Y. Dong, R. von der Heydt, and E. Niebur. Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. *Proceedings of the National Academy of Sciences*, 108(18):7583–8, 2011.

[27] Emine Merve Kaya and Mounya Elhilali. A temporal saliency map for modeling auditory attention. In *Information Sciences and Systems (CISS), 2012 46th Annual Conference on*, pages 1–6. IEEE, 2012.

[28] Christoph Kayser, Rodrigo F Salazar, and Peter König. Responses to natural scenes in cat V1. *Journal of neurophysiology*, 90(3):1910–1920, 2003.

[29] Anatomy of the human eye. `https://www.flickr.com/photos/entirelysubjective/6065758886`, Accessed: 2015-09-30. Licensed type: reuse with modification.

[30] Mark D Fairchild. *Color appearance models*. John Wiley & Sons, 2013.

[31] Haruhisa Okawa and Alapakkam P Sampath. Optimization of single-photon response transmission at the rod-to-rod bipolar synapse. *Physiology*, 22(4):279–286, 2007.

[32] Pathway from retina to visual cortex. `https://en.wikipedia.org/wiki/Visual_system`, Accessed: 2015-10-10. Licensed type: Creative Commons.

[33] Russell L DeValois and Karen K DeValois. *Spatial vision*. Number 14. Oxford University Press, 1988.

[34] Valerio Mante and Matteo Carandini. Mapping of stimulus energy in primary visual cortex. *Journal of neurophysiology*, 94(1):788–798, 2005.

[35] Gregory D Horwitz and Charles A Hass. Nonlinear analysis of macaque v1 color tuning reveals cardinal directions for cortical color processing. *Nature neuroscience*, 15(6):913–919, 2012.

[36] PM Daniel and D Whitteridge. The representation of the visual field on the cerebral cortex in monkeys. *The Journal of physiology*, 159(2):203, 1961.

[37] S. Yantis. *Sensation and Perception*. Worth Publishers, 2013.

[38] Allan Dobbins, Steven W Zucker, and Max S Cynader. Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature*, 329(6138):438–441, 1987.

[39] F. Heitger, L. Rosenthaler, R. von der Heydt, E. Peterhans, and O. Kübler. Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Research*, 32 (5):963–981, 1992.

[40] H. Zhou, H. S. Friedman, and R. von der Heydt. Coding of border ownership in monkey visual cortex. *J. Neurosci.*, 20(17):6594–6611, 2000.

[41] Jay Hegdé and David C Van Essen. Selectivity for complex shapes in primate visual area v2. *The Journal of Neuroscience*, 2000.

[42] Akiyuki Anzai, Xinmiao Peng, and David C Van Essen. Neurons in monkey visual area V2 encode combinations of orientations. *Nature neuroscience*, 10(10):1313–1321, 2007.

[43] Yasmine El-Shamayleh and J Anthony Movshon. Neuronal responses to texture-defined form in macaque visual area v2. *The Journal of neuroscience*, 31(23):8543–8555, 2011.

[44] Ben Willmore, Ryan Prenger, and Jack Gallant. Neural representation of natural images in visual area V2. *The Journal of neuroscience*, 30(6):2102–2114, 2010.

[45] Anitha Pasupathy and Charles E Connor. Responses to contour features in macaque area v4. *Journal of Neurophysiology*, 82(5):2490–2502, 1999.

[46] Anitha Pasupathy and Charles E Connor. Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of neurophysiology*, 86(5):2505–2519, 2001.

[47] Charles G Gross and S De Schonen. Representation of visual stimuli in inferior temporal cortex [and discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 335(1273):3–10, 1992.

[48] Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *The Journal of Neuroscience*, 35(39):13402–13418, 2015.

[49] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[50] Richard T Born and David C Bradley. Structure and function of visual area MT. *Annu. Rev. Neurosci.*, 28:157–189, 2005.

[51] A David Milner and Melvyn A Goodale. Two visual systems re-viewed. *Neuropsychologia*, 46(3):774–785, 2008.

[52] Robert D McIntosh and Thomas Schenk. Two visual streams for perception and action: Current trends. *Neuropsychologia*, 47(6):1391–1396, 2009.

[53] Bosun Xie. *Head-related transfer function and virtual auditory display.* J Ross, 2013.

[54] Anatomy of the human ear. `https://commons.wikimedia.org/wiki/File:Ear-anatomy-text-small-en.svg`, Accessed: 2015-10-11. Licensed type: Creative Commons.

[55] Characteristic frequencies of the basilar membrane. `https://en.wikipedia.org/wiki/Electrocochleography#/media/File:Uncoiled_cochlea_with_basilar_membrane.png`, Accessed: 2017-11-23. Licensed type: Creative Commons.

[56] Manuel S Malmierca and David K Ryugo. Descending connections of auditory cortex to the midbrain and brain stem. In *The auditory cortex*, pages 189–208. Springer, 2011.

[57] JohnE. Mendoza. Trapezoid body. In JeffreyS. Kreutzer, John DeLuca, and Bruce Caplan, editors, *Encyclopedia of Clinical Neuropsychology*, pages 2549–2549. Springer New York, 2011.

[58] Jean K Moore. Organization of the human superior olivary complex. *Microscopy research and technique*, 51(4):403–412, 2000.

[59] Josef P Rauschecker and Sophie K Scott. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, 12(6): 718–724, 2009.

[60] Asif A Ghazanfar and Charles E Schroeder. Is neocortex essentially multisensory? *Trends in cognitive sciences*, 10(6):278–285, 2006.

[61] BE Stein and MA Meredith. Merging of the senses. *MIT Press*, 1993.

[62] Dora E Angelaki, Yong Gu, and Gregory C DeAngelis. Multisensory integration: psychophysics, neurophysiology, and computation. *Current opinion in neurobiology*, 19(4):452–458, 2009.

[63] M Alex Meredith and Barry E Stein. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of neurophysiology*, 56(3):640–662, 1986.

[64] M Alex Meredith and Barry E Stein. Spatial determinants of multisensory integration in cat superior colliculus neurons. *Journal of Neurophysiology*, 75(5):1843–1857, 1996.

[65] M Alex Meredith, James W Nemitz, and Barry E Stein. Determinants of multisensory integration in superior colliculus neurons. I. temporal factors. *The Journal of neuroscience*, 7(10):3215–3229, 1987.

[66] Gopathy Purushothaman, Roan Marion, Keji Li, and Vivien A Casagrande. Gating and control of primary visual cortex by pulvinar. *Nature neuroscience*, 15(6):905–912, 2012.

[67] Rebecca A Berman and Robert H Wurtz. Exploring the pulvinar path to visual cortex. *Progress in brain research*, 171:467–473, 2008.

[68] Holle Kirchner, Emmanuel J Barbeau, Simon J Thorpe, Jean Régis, and Catherine Liégeois-Chauvel. Ultra-rapid sensory responses in the human frontal eye field region. *The Journal of Neuroscience*, 29(23):7599–7606, 2009.

[69] Gemma A Calvert, Charles Spence, and Barry E Stein. *The handbook of multisensory processes*. MIT press, 2004.

[70] Arnaud Falchier, Simon Clavagnier, Pascal Barone, and Henry Kennedy. Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience*, 22(13):5749–5759, 2002.

[71] Kathleen S Rockland and Hisayuki Ojima. Multisensory convergence in calcarine

visual areas in macaque monkey. *International Journal of Psychophysiology*, 50(1): 19–26, 2003.

[72] John F Smiley and Arnaud Falchier. Multisensory connections of monkey auditory cerebral cortex. *Hearing research*, 258(1):37–46, 2009.

[73] Arnaud Falchier, Charles E. Schroeder, Troy A. Hackett, Peter Lakatos, Sheila Nascimento-Silva, Istvan Ulbert, Gyorgi Karmos, and John F. Smiley. Projection from visual areas V2 and prostriata to caudal auditory cortex in the monkey. *Cerebral Cortex*, 20(7):1529–1538, 2010.

[74] James W Lewis and David C Van Essen. Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *Journal of Comparative Neurology*, 428(1):112–137, 2000.

[75] François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.

[76] Chris Edwards. Growing pains for deep learning. *Commun. ACM*, 58(7):14–16, June 2015.

[77] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[78] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled:

High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[79] H. Zhou, H.S. Friedman, and R. Von Der Heydt. Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, 20(17):6594–6611, 2000.

[80] Fangtu T. Qiu and Rdiger von der Heydt. Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron*, 47(1):155 – 166, 2005.

[81] Jonathan R. Williford and Rüdiger von der Heydt. Figure-ground organization in visual cortex for natural scenes. *eNeuro*, 3(6), 2016.

[82] Christopher Chabris and Daniel Simons. *The invisible gorilla: And other ways our intuitions deceive us.* Broadway Books, 2011.

[83] Barry E Stein and Terrence R Stanford. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(4):255–266, 2008.

[84] Katharina Von Kriegstein, Andreas Kleinschmidt, Philipp Sterzer, and Anne-Lise Giraud. Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, 17(3):367–376, 2005.

[85] Sarah Watkins, Ladan Shams, Sachiyo Tanaka, J-D Haynes, and Geraint Rees. Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage*, 31(3):1247–1256, 2006.

[86] Virginie van Wassenhove, Ken W Grant, and David Poeppel. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4):1181–1186, 2005.

[87] Asif A Ghazanfar, Joost X Maier, Kari L Hoffman, and Nikos K Logothetis. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience*, 25(20):5004–5012, 2005.

[88] Ye Wang, Simona Celebrini, Yves Trotter, and Pascal Barone. Visuo-auditory interactions in the primary visual cortex of the behaving monkey: electrophysiological evidence. *BMC neuroscience*, 9(1):79, 2008.

[89] Sudarshan Ramenahalli, Stefan Mihalas, and Ernst Niebur. Local spectral anisotropy is a valid cue for figure–ground organization in natural scenes. *Vision research*, 103: 116–126, 2014.

[90] Sudarshan Ramenahalli, Stefan Mihalas, and Ernst Niebur. Spectral inhomogeneity provides information for figure-ground organization in natural images. *Society for Neuroscience Annual Meeting*, 2011.

[91] Sudarshan Ramenahalli, Stefan Mihalas, and Ernst Niebur. Figure-ground classification based on spectral anisotropy of local image patches. In *Proceedings of the 46th Annual IEEE Conference on Information Sciences and Systems (IEEE-CISS), 2012*, pages 1–5, 2012.

[92] John J McDonald, Wolfgang A Teder-Sälejärvi, and Steven A Hillyard. Involuntary orienting to sound improves visual perception. *Nature*, 407(6806):906–908, 2000.

[93] A. Tsiami, A. Katsamanis, P. Maragos, and A. Vatakis. Towards a behaviorally-validated computational audiovisual saliency model. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851, March 2016.

[94] Sudarshan Ramenahalli, Daniel R Mendat, Salvador Dura-Bernal, Eugenio Culurciello, E Nieburt, and Andreas Andreou. Audio-visual saliency map: overview, basic models and hardware implementation. In *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, pages 1–6. IEEE, 2013.

[95] T. Ghose and S.E. Palmer. Extremal edges versus other principles of figure-ground organization. *Journal of vision*, 10(8), 2010. ISSN 1534-7362.

[96] David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, Dec 1987.

[97] D. J. Tolhurst, Y. Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992.

[98] D.L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994.

[99] Daniel Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994.

[100] A. van der Schaaf and J.H. van Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759 – 2770, 1996.

[101] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1226 – 1238, sep 2002.

[102] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003.

[103] James J. Gibson. *The perception of the visual world*. Oxford, England: Houghton Mifflin, 1950.

[104] Victor Klymenko and Naomi Weisstein. Spatial frequency differences can determine figure-ground organization. *Journal of Experimental Psychology: Human Perception and Performance*, 12(3):324 – 330, 1986.

[105] Johannes Burge, Charless C. Fowlkes, and Martin S. Banks. Natural-scene statistics predict how the figure-ground cue of convexity affects human depth perception. *Journal of Neuroscience*, 30(21):7269–7280, 2010.

[106] Xiaofeng Ren, Charless C Fowlkes, and Jitendra Malik. Figure/ground assignment in natural images. In *European Conference on Computer Vision*, pages 614–627. Springer, 2006.

[107] Wilson S Geisler, Jiri Najemnik, and Almon D Ing. Optimal stimulus encoders for natural tasks. *Journal of vision*, 9(13):17, 2009.

[108] Shohei Matsuoka, Yasuhiro Hatori, and Ko Sakai. Perception of border ownership by multiple gestalt factors. In *Asia Pacific Conference on Vision, 2012*, page 62, Inchon, Korea, July 2012.

[109] Ashutosh Saxena, Andrew Ng, and Sung Chung. Learning Depth from Single Monocular Images. *NIPS*, 18, 2005.

[110] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3D: Learning 3D scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):824–840, May 2009.

[111] Fredric J Harris. On the use of windows for harmonic analysis with the Discrete Fourier Transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

[112] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025, MIT, 2005.

[113] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proc. 8th Int'l Conf. Computer Vision*, 2:416–423, July 2001.

[114] R.C. Gonzalez, R.E. Woods, and S.L. Eddins. *Digital Image Processing Using MATLAB*. Pearson Prentice Hall, 2004. ISBN 9780130085191.

[115] RGB colorspace to grayscale conversion. `http://www.mathworks.com/help/images/ref/rgb2gray.html`, 2014. Accessed: 05-12-2014.

[116] John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, April 2001.

[117] Felix A Wichmann, Jan Drewes, Pedro Rosas, and Karl R Gegenfurtner. Animal

detection in natural scenes: critical features revisited. *Journal of Vision*, 10(4):6, 2010.

[118] Thierry Blu and Michael Unser. Image interpolation and resampling. In *Handbook of Medical Imaging, Processing and Analysis*, pages 393–420. Academic Press, 2000.

[119] J Anthony Parker, Robert V Kenyon, and D Troxel. Comparison of interpolating methods for image resampling. *Medical Imaging, IEEE Transactions on*, 2(1):31–39, 1983.

[120] Ken-Ichiro Tsutsui, Hideo Sakata, Tomoka Naganuma, and Masato Taira. Neural correlates for perception of 3D surface orientation from texture gradient. *Science*, 298 (5592):409–412, 2002.

[121] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

[122] B Schlkopf and A J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, volume 98. MIT Press, 2002.

[123] Chih-wei Hsu, Chih-chung Chang, and Chih-jen Lin. A practical guide to support vector classification. *Bioinformatics*, 1(1):1–16, 2010.

[124] Christopher Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[125] Bryan Russell, Antonio Torralba, and William Freeman. LabelMe: the open annotation tool. URL `http://labelme.csail.mit.edu/`.

[126] V. A. F. Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *J Neurosci*, 15:1605–1615, 1995.

[127] Ko Sakai, Haruka Nishimura, Ryohei Shimizu, and Keiichi Kondo. Consistent and robust determination of border ownership based on asymmetric surrounding contrast. *Neural Networks*, 33:257–274, 2012.

[128] Dario L Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of neurophysiology*, 88(1):455–463, 2002.

[129] Joel Zylberberg, Jason Timothy Murphy, and Michael Robert DeWeese. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields. *PLoS Comput Biol*, 7(10):e1002250, 2011.

[130] Gary A Walker, Izumi Ohzawa, and Ralph D Freeman. Asymmetric suppression outside the classical receptive field of the visual cortex. *The Journal of Neuroscience*, 19(23):10536–10553, 1999.

[131] Dario L Ringach. Mapping receptive fields in primary visual cortex. *The Journal of Physiology*, 558(3):717–728, 2004.

[132] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of Eighth IEEE International Conference on Computer Vision, 2001*, volume 2, pages 416–423. IEEE, 2001.

[133] Hsing-Kuo Pao, Davi Geiger, and Nava Rubin. Measuring convexity for figure/ground separation. *Proceedings of the 7th IEEE International Conference on Computer Vision*, 2:948–955, 1999.

[134] Robert Shapley, Michael Hawken, and Dario L Ringach. Dynamics of orientation selectivity in the primary visual cortex and the importance of cortical inhibition. *Neuron*, 38(5):689–699, 2003.

[135] E. Craft, H. Schutze, E. Niebur, and R. Von Der Heydt. A neural model of figure-ground organization. *Journal of Neurophysiology*, 97(6):4310–4326, 2007.

[136] Pieter R Roelfsema, Victor AF Lamme, Henk Spekreijse, and Holger Bosch. Figure-ground segregation in a recurrent network architecture. *Journal of Cognitive Neuroscience*, 14(4):525–537, 2002.

[137] Li Zhaoping. Border ownership from intracortical interactions in visual area V2. *Neuron*, 47:143–153, 2005.

[138] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11):1254 –1259, November 1998.

[139] F. Heitger and R. von der Heydt. A computational model of neural contour processing: figure-ground segregation and illusory contours. In *Proc. 4th Int. Conf. Computer Vision*, pages 32–40. IEEE Computer Society Press, 1993.

[140] Thorsten Hansen and Heiko Neumann. A biologically motivated scheme for robust

junction detection. *Proceedings of Second International Workshop on Biologically Motivated Computer Vision*, pages 16–26, 2002. doi: 10.1007/3-540-36181-2_2.

[141] Victor AF Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *The Journal of Neuroscience*, 15(2):1605–1615, 1995.

[142] Hans Super and Victor AF Lamme. Altered figure-ground perception in monkeys with an extra-striate lesion. *Neuropsychologia*, 45(14):3329–3334, 2007.

[143] Jonathan R Williford and Rudiger von der Heydt. Early visual cortex assigns border ownership in natural scenes according to image context. *Journal of Vision*, 14(10): 588–588, 2014.

[144] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, 91(3):328–346, 2011.

[145] Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Fast 2D border ownership assignment. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5117–5125, 2015.

[146] Stephen Grossberg and Ennio Mingolla. Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychological review*, 92 (2):173, 1985.

[147] Stephen Grossberg. 3-D vision and figure-ground separation by visual cortex. *Perception & psychophysics*, 55(1):48–121, 1994.

[148] Paul K Kienker, Terrence J Sejnowski, Geoffrey E Hinton, and Lee E Schumacher. Separating figure from ground with a parallel network. *Perception*, 15(2):197–216, 1986.

[149] P Sajda and L.H. Finkel. Intermediate-level visual representations and the construction of surface perception. *J Cogn Neurosci*, 7:267–291, 1995.

[150] Janneke FM Jehee, Victor AF Lamme, and Pieter R Roelfsema. Boundary assignment in a recurrent network architecture. *Vision research*, 47(9):1153–1165, 2007.

[151] Zhaoping Li. V1 mechanisms and some figure–ground and border effects. *Journal of Physiology-Paris*, 97(4):503–515, 2003.

[152] Zhaoping Li. Can V1 mechanisms account for figure-ground and medial axis effects? In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 136–142. MIT Press, 2000.

[153] Mitesh K. Kapadia, Minami Ito, Charles D. Gilbert, and Gerald Westheimer. Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in V1 of alert monkeys. *Neuron*, 15(4):843 – 856, 1995.

[154] Adam M Slllito, Kenneth L Grieve, Helen E Jones, Javier Cudeiro, and Justin Davls. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, 378 (6556):492–496, 1995.

[155] J. J. Knierim and D. C. Van Essen. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophysiology*, 67(4):961–980, 1992.

[156] Masayuki Kikuchi and Youhei Akashi. A model of border-ownership coding in early vision. In *International Conference on Artificial Neural Networks – ICANN*, pages 1069–1074. Springer, 2001.

[157] D. Ardila, S. Mihalas, R. von der Heydt, and E. Niebur. Medial axis generation in a model of perceptual organization. *46th IEEE Annual Conference on Information Sciences and Systems*, pages 1–4, 2012.

[158] Jamal Lottier Molin, Alexander F Russell, Stefan Mihalas, Ernst Niebur, and Ralph Etienne-Cummings. Proto-object based visual saliency model with a motion-sensitive channel. In *Biomedical Circuits and Systems Conference (BioCAS), 2013 IEEE*, pages 25–28, 2013.

[159] Brian Hu and Ernst Niebur. A recurrent neural model for proto-object based contour integration and figure-ground segregation. *Journal of Computational Neuroscience*, Sep 2017. ISSN 1573-6873. doi: 10.1007/s10827-017-0659-3.

[160] Oliver W Layton, Ennio Mingolla, and Arash Yazdanbakhsh. Dynamic coding of border-ownership in visual cortex. *Journal of vision*, 12(13):8, 2012.

[161] Dražen Domijan and Mia Šetić. A feedback model of figure-ground assignment. *Journal of vision*, 8(7):10, 2008.

[162] Hans Super, August Romeo, and Matthias Keil. Feed-forward segmentation of figure-ground and assignment of border-ownership. *PLOS ONE*, 5(5):1–14, 05 2010.

[163] Haruka Nishimura and Ko Sakai. Determination of border ownership based on the surround context of contrast. *Neurocomputing*, 58:843–848, 2004.

[164] Haruka Nishimura and Ko Sakai. The computational model for border-ownership determination consisting of surrounding suppression and facilitation in early vision. *Neurocomputing*, 65:77–83, 2005.

[165] Naoki Kogo, Christoph Strecha, Luc Van Gool, and Johan Wagemans. Surface construction by a 2-D differentiation–integration process: A neurocomputational model for perceived border ownership, depth, and lightness in kanizsa figures. *Psychological review*, 117(2):406, 2010.

[166] Vicky Froyen, Jacob Feldman, and Manish Singh. A bayesian framework for figure-ground interpretation. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 631–639. Curran Associates, Inc., 2010.

[167] Derek Hoiem, Andrew N Stein, Alexei A Efros, and Martial Hebert. Recovering occlusion boundaries from a single image. In *IEEE 11th International Conference on Computer Vision, ICCV, 2007*, pages 1–8, 2007.

[168] Mohamed R. Amer, Raviv Raich, and Sinisa Todorovic. Monocular extraction of 2.1D sketch. In *Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China*, pages 3437–3440, 2010.

[169] Mohamed R. Amer, Siavash Yousefi, Raviv Raich, and Sinisa Todorovic. Monocu-

lar extraction of 2.1D sketch using constrained convex optimization. *International Journal of Computer Vision*, 112(1):23–42, Mar 2015.

[170] Ido Leichter and Michael Lindenbaum. Boundary ownership by lifting to 2.1D. In *IEEE 12th International Conference on Computer Vision, 2009*, pages 9–16. IEEE, 2009.

[171] Guillem Palou and Philippe Salembier. Monocular depth ordering using T-junctions and convexity occlusion cues. *IEEE Transactions on Image Processing*, 22(5):1926–1939, 2013.

[172] Guillem Palou and Philippe Salembier. From local occlusion cues to global monocular depth estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012*, pages 793–796. IEEE, 2012.

[173] Guillem Palou and Philippe Salembier. Occlusion-based depth ordering on monocular images with binary partition tree. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*, pages 1093–1096. IEEE, 2011.

[174] Philippe Salembier and Luis Garrido. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE Transactions on Image Processing*, 9(4):561–576, 2000.

[175] Morimichi Nishigaki, Cornelia Fermüller, and Daniel DeMenthon. The image torque operator: A new tool for mid-level vision. In *IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA*, pages 502–509, 2012.

[176] Stella X. Yu, Tai Sing Lee, and Takeo Kanade. A hierarchical markov random field model for figure-ground segregation. *Proceedings of Third International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 118–133, 2001.

[177] Kyungim Baek and Paul Sajda. Inferring figure-ground using a recurrent integrate-and-fire neural circuit. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(2):125–130, 2005.

[178] Michael Maire. Simultaneous segmentation and figure/ground organization using angular embedding. In *European Conference on Computer Vision–ECCV*, pages 450–464. Springer, 2010.

[179] SX Yu. Angular embedding: from jarring intensity differences to perceived luminance. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009.*, pages 2302–2309. IEEE, 2009.

[180] Adrian Ion, Joao Carreira, and Cristian Sminchisescu. Image segmentation by figure-ground composition into maximal cliques. In *IEEE International Conference on Computer Vision*, pages 2110–2117. IEEE, 2011.

[181] Adrian Ion, João Carreira, and Cristian Sminchisescu. Probabilistic joint image segmentation and labeling by figure-ground composition. *International Journal of Computer Vision*, 107(1):40–57, 2014.

[182] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A.*, 2:284–299, 1985.

[183] MATLAB. 2-D cross-correlation. `https://www.mathworks.com/help/signal/ref/xcorr2.html`, Accessed: 2013-09-30.

[184] Eric W. Weisstein. von Mises Distribution. `http://mathworld.wolfram.com/vonMisesDistribution.html`, Accessed: 2014-09-30.

[185] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, May 2011.

[186] Marius Leordeanu, Rahul Sukthankar, and Cristian Sminchisescu. Generalized boundaries from multiple image interpretations. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1312–1324, 2014.

[187] George Azzopardi, Antonio Rodríguez-Sánchez, Justus Piater, and Nicolai Petkov. A push-pull CORF model of a simple cell with antiphase inhibition improves SNR and contour detection. *PLoS One*, 9(7):e98424, 2014.

[188] Lee A. Iverson and Steven W. Zucker. Logical/linear operators for image curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):982–996, 1995.

[189] Peter Ulric Tse and Marc K Albert. Amodal completion in the absence of image tangent discontinuities. *Perception*, 27(4):455–464, 1998.

[190] Josh McDermott. Psychophysics with junctions in real images. *Perception*, 33(9): 1101–1127, 2004.

[191] Peter A van der Helm. Bayesian confusions surrounding simplicity and likelihood in perceptual organization. *Acta psychologica*, 138(3):337–346, 2011.

[192] Naoki Kogo and Raymond van Ee. Neural mechanisms of figure-ground organization: Border-ownership, competition and perceptual switching. *Handbook of perceptual organization, Oxford University Press UK*, 2015.

[193] P.S. Huggins, H.F. Chen, P.N. Belhumeur, and S.W. Zucker. Finding folds: On the appearance and identification of occlusion. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–718. IEEE, 2001.

[194] S.E. Palmer and T. Ghose. Extremal Edge– A Powerful Cue to Depth Perception and Figure-Ground Organization. *Psychological Science*, 19(1):77, 2008. ISSN 0956-7976.

[195] S. Ramenahalli, S. Mihalas, and E. Niebur. Extremal edges: Evidence in natural images. In *45th Annual Conference on Information Sciences and Systems (CISS), 2011*, pages 1 –5, march 2011.

[196] A. F. Russell, S Mihalas, R. von der Heydt, E. Niebur, and R. Etienne-Cummings. A model of proto-object based saliency. *Vision Research*, 94:1–15, 2014.

[197] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature Publishing Group*, 1976.

[198] David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262, 2004.

[199] Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. Visual illusion induced by sound. *Cognitive Brain Research*, 14(1):147–152, 2002.

[200] David Alais, Fiona N Newell, and Pascal Mamassian. Multisensory processing in review: from physiology to behaviour. *Seeing and perceiving*, 23(1):3–38, 2010.

[201] AJ King and AR Palmer. Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research*, 60(3): 492–500, 1985.

[202] Charles Spence, Daniel Senkowski, and Brigitte Röder. Crossmodal processing. *Experimental Brain Research*, 198(2):107–111, 2009.

[203] Georgios Evangelopoulos, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, A Zlatintsi, and Yair Avrithis. Movie summarization based on audiovisual saliency detection. In *15th IEEE International Conference on Image Processing*, pages 2528–2531. IEEE, 2008.

[204] Guanghan Song. *Effet du son dans les vidéos sur la direction du regard: contribution à la modélisation de la saillance audiovisuelle.* PhD thesis, Université de Grenoble, 2013.

[205] Stephen Grossberg, Karen Roberts, Mario Aguilar, and Daniel Bullock. A neural model of multimodal adaptive saccadic eye movement control by superior colliculus. *The Journal of neuroscience*, 17(24):9706–9725, 1997.

[206] Matthew C Casey, Athanasios Pavlou, and A Timotheou. Audio-visual localization

with hierarchical topographic maps: Modeling the superior colliculus. *Neurocomputing*, 97:344–356, 2012.

[207] Juan Huo and Alan Murray. The adaptation of visual and auditory integration in the barn owl superior colliculus with spike timing dependent plasticity. *Neural Networks*, 22(7):913–921, 2009.

[208] Juan Huo, Alan Murray, and Dongqing Wei. Adaptive visual and auditory map alignment in barn owl superior colliculus and its neuromorphic implementation. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(9):1486–1497, 2012.

[209] Thomas J Anastasio, Paul E Patton, and Kamel Belkacem-Boussaid. Using bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, 12(5):1165–1187, 2000.

[210] Paul Patton, Kamel Belkacem-Boussaid, and Thomas J Anastasio. Multimodality in the superior colliculus: an information theoretic analysis. *Cognitive Brain Research*, 14(1):10–19, 2002.

[211] Paul E Patton and Thomas J Anastasio. Modeling cross-modal enhancement and modality-specific suppression in multisensory neurons. *Neural computation*, 15(4): 783–810, 2003.

[212] Hans Colonius and Adele Diederich. Why arent all deep superior colliculus neurons multisensory? a bayes ratio analysis. *Cognitive, Affective, & Behavioral Neuroscience*, 4(3):344–353, 2004.

[213] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.

[214] Kevin Wilson, Vibhav Rangarajan, Neal Checka, and Trevor Darrell. Audiovisual arrays for untethered spoken interfaces. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, pages 389–, Washington, DC, USA, 2002. IEEE Computer Society.

[215] Freddy Torres and Hari Kalva. Influence of audio triggered emotional attention on video perception. In *IS&T/SPIE Electronic Imaging*, pages 901408–901408. International Society for Optics and Photonics, 2014.

[216] Jong-Seok Lee, Francesca De Simone, and Touradj Ebrahimi. Efficient video coding based on audio-visual focus of attention. *Journal of Visual Communication and Image Representation*, 22(8):704–711, 2011.

[217] Martin Rerabek, Hiromi Nemoto, Jong-Seok Lee, and Touradj Ebrahimi. Audiovisual focus of attention and its application to ultra high definition video compression. In *IS&T/SPIE Electronic Imaging*, pages 901407–901407. International Society for Optics and Photonics, 2014.

[218] Jonas Ruesch, Manuel Lopes, Alexandre Bernardino, Jonas Hornstein, José Santos-Victor, and Rolf Pfeifer. Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In *IEEE International Conference on Robotics and Automation*, pages 962–967. IEEE, 2008.

[219] Boris Schauerte, Jan Richarz, Thomas Plötz, Christian Thurau, and Gernot A Fink. Multi-modal and multi-camera attention in smart environments. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 261–268. ACM, 2009.

[220] Boris Schauerte, Benjamin Kuhn, Kristian Kroschel, and Rainer Stiefelhagen. Multimodal saliency-based attention for object-based scene analysis. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1173–1179. IEEE, 2011.

[221] Boris Schauerte. Bottom-up audio-visual attention for scene exploration. In *Multimodal Computational Attention for Scene Understanding and Robotics*, pages 35–113. Springer, 2016.

[222] Selim Onat, Klaus Libertus, and Peter Konig. Integrating audiovisual information for the control of overt attention. In *Journal of Vision, 7(10):11.* 2007.

[223] B Kühn, B Schauerte, R Stiefelhagen, and K Kroschel. A modular audio-visual scene analysis and attention system for humanoid robots. In *Proc. 43rd Int. Symp. Robotics (ISR)*, 2012.

[224] Benjamin Kühn, Boris Schauerte, Kristian Kroschel, and Rainer Stiefelhagen. Multimodal saliency-based attention: A lazy robot's approach. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 807–814. IEEE, 2012.

[225] Johannes Bauer, Cornelius Weber, and Stefan Wermter. A som-based model for multi-sensory integration in the superior colliculus. In *The 2012 international joint conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2012.

[226] Raquel Viciana-Abad, Rebeca Marfil, Jose M Perez-Lorenzo, Juan P Bandera, Adrian Romero-Garces, and Pedro Reche-Lopez. Audio-visual perception system for a humanoid robotic head. *Sensors*, 14(6):9522–9545, 2014.

[227] Georgios Evangelopoulos, Konstantinos Rapantzikos, Petros Maragos, Yannis Avrithis, and Alexandros Potamianos. Audiovisual attention modeling and salient event detection. In *Multimodal Processing and Interaction*, pages 1–21. Springer, 2008.

[228] Konstantinos Rapantzikos, Georgios Evangelopoulos, Petros Maragos, and Yannis Avrithis. An audio-visual saliency model for movie summarization. In *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pages 320–323. IEEE, 2007.

[229] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *Multimedia, IEEE Transactions on*, 15(7):1553–1568, 2013.

[230] Jiro Nakajima, Akihiro Sugimoto, and Kazuhiko Kawamoto. Incorporating audio signals into constructing a visual saliency map. In *Image and Video Technology*, pages 468–480. Springer, 2014.

[231] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems (NIPS 2005)*, volume 19, pages 1–8. 2006.

[232] Jiro Nakajima, Akisato Kimura, Akihiro Sugimoto, and Kunio Kashino. Visual attention driven by auditory cues. In *MultiMedia Modeling*, pages 74–86. Springer, 2015.

[233] Danil Korchagin, Petr Motlicek, Stefan Duffner, and Hervé Bourlard. Just-in-time multimodal association and fusion from home entertainment. In *2011 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–5. IEEE, 2011.

[234] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems*, pages 813–819, 2000.

[235] Dante A Blauth, Vicente P Minotto, Claudio R Jung, Bowon Lee, and Ton Kalker. Voice activity detection and speaker localization using audiovisual cues. *Pattern Recognition Letters*, 33(4):373–380, 2012.

[236] Rémi Ratajczak, Denis Pellerin, Quentin Labourey, and Catherine Garbay. A fast audiovisual attention model for human detection and localization on a companion robot. In *The First International Conference on Applications and Systems of Visual Paradigms (VISUAL 2016)*, 2016.

[237] Guanghan Song, Denis Pellerin, and Lionel Granjon. How different kinds of sound in videos can influence gaze. In *13th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE, 2012.

[238] Antoine Coutrot and Nathalie Guyader. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision*, 14(8):5, 2014.

[239] Antoine Coutrot and Nathalie Guyader. An audiovisual attention model for natural conversation scenes. In *IEEE International Conference on Image Processing (ICIP)*, pages 1100–1104. IEEE, 2014.

[240] Athanasios Noulas, Gwenn Englebienne, and Ben JA Kröse. Multimodal speaker diarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1): 79–93, 2012.

[241] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.

[242] Naty Ould Sidaty, Mohamed-Chaker Larabi, and Abdelhakim Saadane. Towards understanding and modeling audiovisual saliency based on talking faces. In *Tenth International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, pages 508–515. IEEE, 2014.

[243] Ryan A Stevenson and Thomas W James. Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage*, 44(3):1210–1223, 2009.

[244] Cliodhna Quigley, Selim Onat, Sue Harding, Martin Cooke, and Peter König. Audio-visual integration during overt visual attention. *Journal of Eye Movement Research*, 1(2):1–17, 2008.

[245] Jingjing Yang, Qi Li, Yulin Gao, and Jinglong Wu. Task-irrelevant auditory stimuli affect audiovisual integration in a visual attention task: Evidence from event-related

potentials. In *IEEE/ICME International Conference on Complex Medical Engineering (CME)*, pages 248 –253, may 2011.

[246] HE Çetingül, Engin Erzin, Yucel Yemez, and A Murat Tekalp. Multimodal speaker/speech recognition using lip motion, lip texture and audio. *Signal processing*, 86(12):3549–3558, 2006.

[247] Satoshi Tamura, Koji Iwano, and Sadaoki Furui. Toward robust multimodal speech recognition. In *Symposium on Large Scale Knowledge Resources (LKR2005)*, pages 163–166, 2005.

[248] Ruth Kimchi, Yaffa Yeshurun, and Aliza Cohen-Savransky. Automatic, stimulus-driven attentional capture by objecthood. *Psychonomic Bulletin & Review*, 14(1): 166–172, 2007.

[249] Antje Nuthmann and John M Henderson. Object-based attentional selection in scene viewing. *Journal of vision*, 10(8):20–20, 2010.

[250] Ronald A Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3): 17–42, 2000.

[251] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2439. IEEE, 2010.

[252] Dequing Sun, Stefan Roth, and Michael Black. Optic flow estimation matlab code. `http://cs.brown.edu/~dqsun/research/software.html`, 2008.

[253] Adam O Donovan, Ramani Duraiswami, and Jan Neumann. Microphone arrays as generalized cameras for integrated audio visual processing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[254] Jens Meyer and Gary Elko. A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages II–1781. IEEE, 2002.

[255] Adam O'Donovan, Ramani Duraiswami, Nail Gumerov, et al. Real time capture of audio images and their use with video. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 10–13. IEEE, 2007.

[256] N.R. Zhang and R. von der Heydt. Analysis of the context integration mechanisms underlying figure–ground organization in the visual cortex. *The Journal of Neuroscience*, 30(19):6482–6496, 2010.

[257] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence*, 34 (11):2189–2202, 2012.

[258] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[259] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards

real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[260] M. I. Posner and Y. Cohen. Components of visual orienting. In H. Bouma and D. G. Bouwhuis, editors, *Attention and Performance X*, pages 531–556. Hilldale, NJ, 1984.

[261] N Van der Stoep, S Van der Stigchel, TCW Nijboer, and C Spence. Visually induced inhibition of return affects the integration of auditory and visual information. *Perception*, 46(1):6–17, 2017.

[262] Charles Spence and Jon Driver. Auditory and audiovisual inhibition of return. *Attention, Perception, & Psychophysics*, 60(1):125–139, 1998.

[263] Tom Troscianko, Rachel Montagnon, Jacques Le Clerc, Emmanuelle Malbert, and Pierre-Louis Chanteau. The role of colour as a monocular depth cue. *Vision Research*, 31(11):1923 – 1929, 1991.

[264] Qasim Zaidi and Andrea Li. Three-dimensional shape perception from chromatic orientation flows. *Visual Neuroscience*, 23(3-4):323330, 2006.

[265] D. Ardila, S. Mihalas, and E. Niebur. How perceptual grouping affects the salience of symmetry. In *Annual Meeting*, page Abstract 801.01/LL18, Washington DC, 2011. Society for Neuroscience.

[266] Stephen E Palmer and Joseph L Brooks. Edge-region grouping in figure-ground organization and depth perception. *Journal of Experimental Psychology: Human Perception and Performance*, 34(6):1353, 2008.

[267] Elan Barenholtz and Michael J. Tarr. Figureground assignment to a translating contour: A preference for advancing vs. receding motion. *Journal of Vision*, 9(5):27, 2009.

[268] Janice E Huss and James A Pennline. A comparison of five benchmarks. *NASA Technical Memorandum 88956*, 1987.

[269] Alexander M Davie and Andrew James Stothers. Improved bound for complexity of matrix multiplication. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 143(2):351–369, 2013.

[270] Narayanan Sundaram. *Making computer vision computationally efficient*. University of California, Berkeley, 2012.

Vita

Sudarshan Ramenahalli

Ph.D. Candidate, Electrical and Computer Engineering,

Johns Hopkins University, Baltimore, MD

email: sramena1@jhu.edu

Thesis

Title: Computational Models of Perceptual Organization and Bottom-up Attention in Visual and Audio-Visual Environments

Advisor: Prof. Ralph Etienne-Cummings

Committee: Prof. Hynek Hermansky, Prof. Mounya Elhilali

Education

Ph.D., Electrical and Computer Engineering (Expected, 2017)

Whiting School of Engineering, Johns Hopkins University, Baltimore, MD

M.S.E, Electrical and Computer Engineering (May, 2011)

Whiting School of Engineering, Johns Hopkins University, Baltimore, MD

M.E.E, Electrical Engineering (May, 2009)

Samuel Ginn College of Engineering, Auburn University, Auburn, AL

B.S., Electronics and Communications (June, 2004)

Sri Jayachamarajendra Coll. of Engg, Visveswaraya Tech Univ, Mysore, India

Selected Research and Work Experience

Visiting Research Scientist, UC Berkeley, Berkeley, CA

Computer Vision R&D Engineer, Iris Automation, San Francisco, CA

Graduate Research Assistant, Johns Hopkins University, Baltimore, MD

Graduate Research Assistant, Auburn University, Auburn, AL

Senior Software Engineer, Robert Bosch Corp, Bangalore, India

Publications

1. Sudarshan Ramenahalli, Stefan Mihalas, Ernst Niebur, "Local spectral anisotropy is a valid cue for figure-ground organization", Vision Research, Vol 103, pp 116 – 126, October 2014
2. Sudarshan Ramenahalli, Ernst Niebur, Ralph Etienne-Cummings, "A model of figure ground organization incorporating local and global cues," (to be submitted to a journal)
3. Sudarshan Ramenahalli, Ernst Niebur, Ralph Etienne-Cummings, "A proto-

object based audiovisual saliency map," (to be submitted to a conference)

4. Daniel Mendat, James E. West, Sudarshan Ramenahalli, Ernst Niebur, Sudarshan Ramenahalli, "Audio-Visual Beamforming with the Eigenmike Microphone Array Omni-Camera and Cognitive Auditory Features", IEEE Conference on Information Sciences and Systems, March 2017
5. Daniel Mendat, James E. West, Sudarshan Ramenahalli, Ernst Niebur, Sudarshan Ramenahalli, "Perceptual Signal Processing for Audio-Visual Beamforming with the Eigenmike Microphone Array and an Omni-Camera", IEEE International Symposium on Circuits and Systems, 2017
6. Sudarshan Ramenahalli, Daniel Mendat, Salvador Dura-Bernal, Eugenio Culurciello, Ernst Niebur, Andreas Andreou, "Audio-visual saliency map: Overview, basic models and hardware implementation," Proceedings of the 47th Annual Conference on Information Sciences and Systems, 2013
7. Sudarshan Ramenahalli, Ernst Niebur, "Computing 3D saliency from a 2D image," Proceedings of the 47th Annual Conference on Information Sciences and Systems, 2013
8. Lalor, N. Mesgarani, S. Rajaram, A. O'Donovan, J. Wright, I. Choi, J. Brumberg, N. Ding, K. C. Lee, N. Peters, S. Ramenahalli, J. Pompe, B. Shinn-Cunningham, M. Slaney, S. Shamma, "Decoding Auditory Attention (in Real Time) with EEG," Association for Research in Otolaryngology (ARO) 36th MidWinter Meeting, February 16-20, 2013.

9. Sudarshan Ramenahalli, Stefan Mihalas, Ernst Niebur, "Figure-ground classification based on spectral properties of boundary image patches," Proceedings of the 46th Annual Conference on Information Sciences and Systems, 2012
10. Sudarshan Ramenahalli, Stefan Mihalas, Ernst Niebur, "Spectral heterogeneity provides information for figure-ground organization in natural images," Society for Neuroscience Annual Meeting 2011
11. Sudarshan Ramenahalli, Stefan Mihalas, Ernst Niebur, "Extremal Edges: Evidence in Natural Images," Proceedings of the 44th Annual Conference on Information Sciences and Systems, 2011
12. X. Liu, S. Ramenahalli, H. Shinagawa, M. Stone, J. L. Prince, E. Murano, J. Zhuo, R. Gullapalli, "Tracking Muscle Deformation During Speech from Tagged and Diffusion Tensor MRI," Joint 159th Acoustical Society of America Meeting 2010
13. Sudarshan Ramenahalli, Thomas Denney, "3D visualization of cardiac tagged magnetic resonance image data using Non-Uniform Rational B-Splines," Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing, 2010

Honors and Awards

- Dean's Fellowship, Johns Hopkins University, Baltimore, MD
- Graduate tuition fellowship, Auburn University, Auburn, AL

- Awarded scholarship to attended Telluride Neuromorphic Cognition Engineering workshop, 2012
- Invited reviewer for 2015 International Conference on Fuzzy System and Data Mining (FSDM2015)
- Recipient of National Talent Search Examination (NTSE) scholarship with 32nd rank (India)

Skills

Python, R, Matlab, Caffe, Torch, Theano, TensorFlow, MatConvNet, C

C++, OpenCV, LaTeX, MS Office, Linux/UNIX, Adobe CS4, InkScape